VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

NÁSTROJ PRO TESTOVÁNÍ KVALITY VÍCENÁSOBNÉHO ZAROVNÁNÍ BIOLOGICKÝCH SEKVENCÍ
A TOOL FOR ASSESSING THE QUALITY OF MULTIPLE SEQUENCE ALIGNMENTS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Veronika Pelikánová**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Helena Škutková, Ph.D.**
SUPERVISOR

BRNO 2016

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Biomedicínská technika a bioinformatika**
Ústav biomedicínského inženýrství

Studentka: Veronika Pelikánová ***ID:*** 164991
Ročník: 3 ***Akademický rok:*** 2015/16

NÁZEV TÉMATU:

Nástroj pro testování kvality vícenásobného zarovnání biologických sekvencí

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši veřejně dostupných nástrojů pro vícenásobné zarovnání biologických sekvencí. 2) Vypracujte literární rešerši parametrů pro popis kvality mnohonásobného zarovnání. 3) Navrhněte metodiku testování kvality zarovnání sekvencí na základě biologického i informačního obsahu sekvencí a dílčí části realizujte v Matlabu. 4) Vytvořte nástroj s GUI v Matlabu pro vyhodnocení kvality zarovnání biologických sekvencí. 5) Vytvořte soubor nukleotidových a aminokyselinových sekvencí pro testování kvality zarovnání z veřejně dostupných databází. 6) Vytvořený nástroj použijte na srovnání kvality zarovnání veřejně dostupných online nástrojů na připravených setech sekvencí.

DOPORUČENÁ LITERATURA:

[1] THOMPSON, JULIE D., BENJAMIN LINARD, ODILE LECOMPTE a OLIVIER POCH. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. PLoS ONE, 2011, 6(3), e18093.

[2] AHOLA, V., T. AITTOKALLIO, M. VIHINEN a E. UUSIPAIKKA. A statistical score for assessing the quality of multiple sequence alignments. BMC Bioinformatics, 2006, 7, 484.

Termín zadání: 8.2.2016 ***Termín odevzdání:*** 27.5.2016

Vedoucí práce: Ing. Helena Škutková, Ph.D.
Konzultant bakalářské práce:

prof. Ing. Ivo Provazník, Ph.D., *předseda oborové rady*

Abstrakt:

Tato bakalářská práce se zabývá testováním kvality vícenásobného zarovnání biologických sekvencí. Obsahuje literární rešerši veřejně dostupných nástrojů vícenásobného zarovnání. Stěžejní bod práce tkví v programu pro testování kvality vícenásobného zarovnání (vytvořené pomocí různých nástrojů, při odlišném nastavení, aj.). Program má uživatelské rozhraní vytvořené v Matlabu a je k práci přiložen společně se souborem nukleotidových a proteinových sekvencí, sestavených za účelem testování a aplikování programu. V neposlední řadě je v práci obsaženo hodnocení veřejně dostupných nástrojů vytvořené na základě výpočtů programu.

Klíčová slova:

Vícenásobné zarovnání, vodící strom, penalizace mezer, Z-skóre, entropie, suma párů, sloupcové skóre

Abstract:

This thesis is dealing with quality assessment of multiple sequence alignments. It contains literature review of public available tools of multiple sequence alignments. The main part of this thesis focuses on the program on quality assessment of multiple sequence alignments (created with the help of various tools, using different settings, etc.). The program has a graphical user interface created in Matlab and is enclosed together with a set of nucleotide and protein sequences compiled for program testing and aplication. Last but not least, the thesis contains the assessment of public available tools resulting from program outputs.

Keywords:

Multiple sequence alignment, guide tree, gap penalty, Z-score, entropie, Sum of pair, column score

PELIKÁNOVÁ, V. *Nástroj pro testování kvality vícenásobného zarovnání biologických sekvencí*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2016. 69 s. Vedoucí bakalářské práce Ing. Helena Škutková, Ph.D..

Prohlášení autora o původnosti díla:

Prohlašuji, že svou bakalářskou práci na téma Nástroj pro testování kvality vícenásobného zarovnání biologických sekvencí jsem vypracovala samostatně pod vedením vedoucí bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujícího zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů, ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne 25.5.2016 ……

Podpis autora

Poděkování:

Moc děkuji vedoucí bakalářské práce Ing. Heleně Škutkové, Ph.D. za její cenné rady, odbornou pomoc, skvělé vedení. Velice obdivuji její pružné reakce. Dále panu doc. RNDr. Jaromíru Baštincovi, CSc. za rady ohledně statistického hodnocení výsledků. V neposlední řadě si velký dík zaslouží rodina, která mě při psaní bakalářské práce a celém studiu podporovala.

Obsah

Úvod

Genetika hraje v posledních letech stále větší roli ve výzkumu i v lékařství (genová terapie). Geny jsou mapovány, zjišťuje se jejich podobnost s různými organismy, porovnávají se zdravé a poškozené geny atd. K tomu aby se vůbec geny mohly hodnotit, porovnávat je tu bioinformatika.

Vícenásobné zarovnání (zarovnání více než dvou sekvencí) je v bioinformatice jedním ze základních nástrojů. Jeho hlavním úkolem je nalézt shodné oblasti v předložených sekvencích, tyto oblasti zarovnat pod sebe, tzn. vzájemně pozičně přizpůsobit. Místa, nacházející se mezi těmito důležitými oblastmi, jsou doplněny mezerami. Problém zarovnání je velmi komplexní záležitost, protože je třeba brát v potaz biologické vlastnosti sekvencí, strukturní zohlednění, funkčnost. Špatné zarovnání by např. mohlo ignorovat funkci, strukturu či příbuznost sekvencí nevhodným vložením mezer. Správné vícenásobné zarovnání umožňuje získat další přidanou hodnotu. Používá se například při vytváření konsenzuálních sekvencí, predikcí funkcí genů, určování struktury proteinu, fylogenetických rekonstrukcích, analýzách proteinových rodin, vyhledávání příbuzných motivů, atd.

K tomuto účelu je vytvořena řada nástrojů. Ty se mohou lišit použitými postupy, nastaveními, preferencemi, aj. Cílem práce je vytvořit program, který dokáže ohodnotit zarovnání (různého nastavení, z odlišných veřejně dostupných nástrojů) a na základě vybraných parametrů doporučí to nejvhodnější. Parametry by měly zohledňovat biologický i informační obsah. Biologický obsah je zásadní pro dodržení přírodních zákonů přepisu genů, funkčnost. Informační obsah je zastoupen entropií. Pomocí ní se snažíme pochopit strukturu zápisu sekvencí.

V bakalářské práci jsou veřejně dostupné algoritmy pro vícenásobné zarovnání testovány pomocí vybraných sekvencí z databáze BaliBase a z RNA porovnávacího webu, které jsou za účelem testování vícenásobného zarovnání sestaveny. Obsahují také ručně vytvořená zarovnání, která jsou stále srovnávaným standardem programovému zarovnání. Tyto sady jsou pomocí veřejně dostupných nástrojů zarovnány a podle různých parametrů ohodnoceny. Výstupem je program pro hodnocení kvality vícenásobného zarovnání a doporučení pro použití veřejně dostupných nástrojů pro daný typ sekvencí lišící se např. podobností sekvencí, délkami sekvencí, uspořádáním.

Daná problematika bohužel není dostatečně popsána, to byl jeden z důvodů k vytvoření této práce. V genetice je otázka vícenásobného zarovnání celkem zásadní a je třeba zvolit nástroj, který se pro konkrétní potřebu hodí nejlépe.

1 Nástroje vícenásobného zarovnání

Ve veškerých technologiích dochází k neustálému rozvoji, zrychlování, posunování hranic o kousek dále. Tak je tomu i v bioinformatice. Proto bylo a je nutné vyvíjet nové nástroje, které budou využívat nejnovější postupy, či přizpůsobovat, zlepšovat stávající nástroje.

K vícenásobnému zarovnání jsou různé přístupy. Existuje mnoho veřejně dostupných nástrojů pro toto zarovnání. Do této práce jich bylo vybráno 7 nejčastěji používaných. Inspirací při jejich výběru byly i přehledové studie [22] a [33]. Jsou to nástroje Clustal, T-Coffee, Mafft, Muscle, Dialign, Kalign, ProbCons a funkce *multialign* z bioinformatického toolboxu programového prostředí Matlab. Jelikož jsou použity různé algoritmy a parametry, zarovnání stejných sekvencí může být dost odlišné. Je potřeba nástroje znát a pro danou sekvenci vybrat ten vhodný.

1.1 Nástroj Clustal

Clustal je progresivní metoda pro vícenásobné zarovnání různých biologických sekvencí. Byl uveden v roce 1988 pány Desmond G. Higgins a Paul M. Sharp z Trinity College z Dublinu, patří k nejstarším nástrojům pro zarovnání vůbec. Algoritmus této nejstarší metody vícenásobného zarovnání je uvedena v Příloze I. [1]

Clustal se postupně vyvíjel. Jeho první aktualizace přišla roku 1991. Verze byla nazvána ClustalV. Původní Clustal byl přepsán a doplněn o funkce, které umožnily nově ukládat a znovu použít již zarovnané sekvence. Bylo možné je doplnit o další sekvence a aktualizovat starý genetický strom. [2][3]

Třetí generace Clustal byla uvedena roku 1994 s označením ClustalemW. ClustalW byl doplněn ClustalemX, roku 1997 [4]. Nejnovější verze je ovšem ClustalOmega. Dále v textu jsou tyto algoritmy popsány. [3]

ClustalW a ClustalX

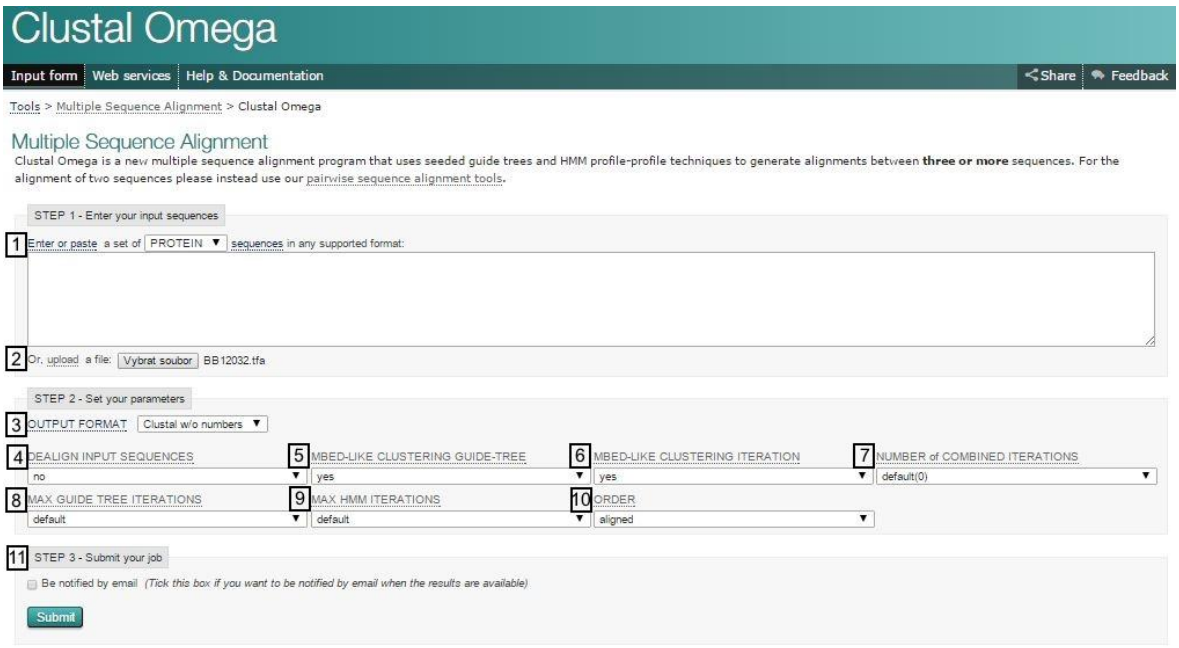
ClustalW byl v říjnu 2015 úplně nahrazen Clustalem Omega. Přestože již není veřejně dostupný, představení je namístě, protože byl řadu let jedním z nejpoužívanějších.

Zarovnání pomocí nástroje ClustalW probíhalo následujícím způsobem. Nejprve byly zarovnány dvojice sekvencí, zjistily se jednotlivé vzdálenosti mezi páry a byla vytvořena matice distancí. Poté se sestavil vodící strom (z anglického guide tree) pomocí Neighbour-Joining methody (dále. NJ metody, na rozdíl od Clustal, kde se využívala metoda UPGMA). Z vodícího stromu byly vybrány nejpodobnější sekvence a ty se zarovnaly, poté k nim byly přirovnávány další nejbližší sekvence, či zarovnané sobě blízké sekvence, až došlo k zarovnání všech sekvencí. [6]

ClustalX je rozšířený ClustalW o uživatelské rozhraní.

Clustal Omega

Clustal Omega je nejnovější verze Clustalu. V algoritmu se projevuje snaha o zrychlení procesu zarovnání. Pro vytvoření vodícího stromu je použit tzv. „emBedding“ (není použita celá sekvence, ale pouze její část). To umožňuje veliké zrychlení a následné vytvoření shluků, pomocí metody k-means, či UPGMA. Samotné zarovnání probíhá využitím tzv. HHalign, jež má mechanismus založen na skrytých Markovových modelech (HMM, viz. [8]). [7] Dále v hodnocení je použita právě tato verze. Pro zjednodušení je Clustal Omega uváděn pouze pod názvem Clustal.



Obr. 1: Čelní panel online nástroje Clustal Omega, výchozí nastavení.

Při práci s nástrojem Clustal Omega je možné měnit následující parametry a tím ovlivnit výsledek zarovnání. Čísla viz. Obr. 1.

1. Volba typu sekvencí Protein/DNA/RNA
2. Vstupní sekvence je možné vkládat ve formátech GCG, FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP nebo UniProtKB/Swiss-Prot, velikost omezena na 2000 sekvencí a 2 MB
3. Výstupní formát zarovnání: Clustal w/o numbers, Clustal w/ numbers, Pearson/FASTA, MSF, NEXUS, PHYLIP,SELEX,STROCKHOLM, VIENNA
4. Možnosti ano/ne. Ano: před spuštěním zarovnání jsou ze sekvencí odstraněny mezery.
5. Možnosti ano/ne. Ano: použití tzv. „emBeding“ k vytvoření vodícího stromu (jsou použity pouze části sekvencí) dochází ke zrychlení algoritmu, zmenšení výpočetní náročnosti, naopak může dojít k nepřesnostem.
6. Možnosti ano/ne. Ano: použití tzv. „emBeding“ v iteračních procesech

7. Tímto parametrem ovlivňujeme počet iterací HMM s následným přepočtem vodícího stromu. Původně je nastavena 0, je možné nastavit až hodnotu 5. Tento parametr je vhodné měnit až v případě, kdy zarovnáváme 1000 sekvencí či více. Při malém počtu může mít na zarovnání naopak iterace špatný vliv.
8. Tento parametr je možné měnit, pouze pokud máme v parametru 7 nastaveny nějaké iterace. Pokud chceme mít počet iterací HMM a vodícího stromu odlišný, můžeme to vyjádřit zde tím, že omezíme maximální počet iterací vodícího stromu. (možnost nastavení 0-5).
9. Platí to samé jako u parametru 8 s rozdílem omezení maximálního počtu HMM iterací.
10. Možnosti: seřadit (aligned) / vstup (input). Pro možnost seřadit je ve výstupním zarovnání řazení sekvencí pozměněno. Pro možnost vstup je ponecháno pořadí sekvencí v zarovnání shodné, jako pořadí vstupních sekvencí.
11. V případě zájmu, je možné nechat si odeslat výsledky na email.

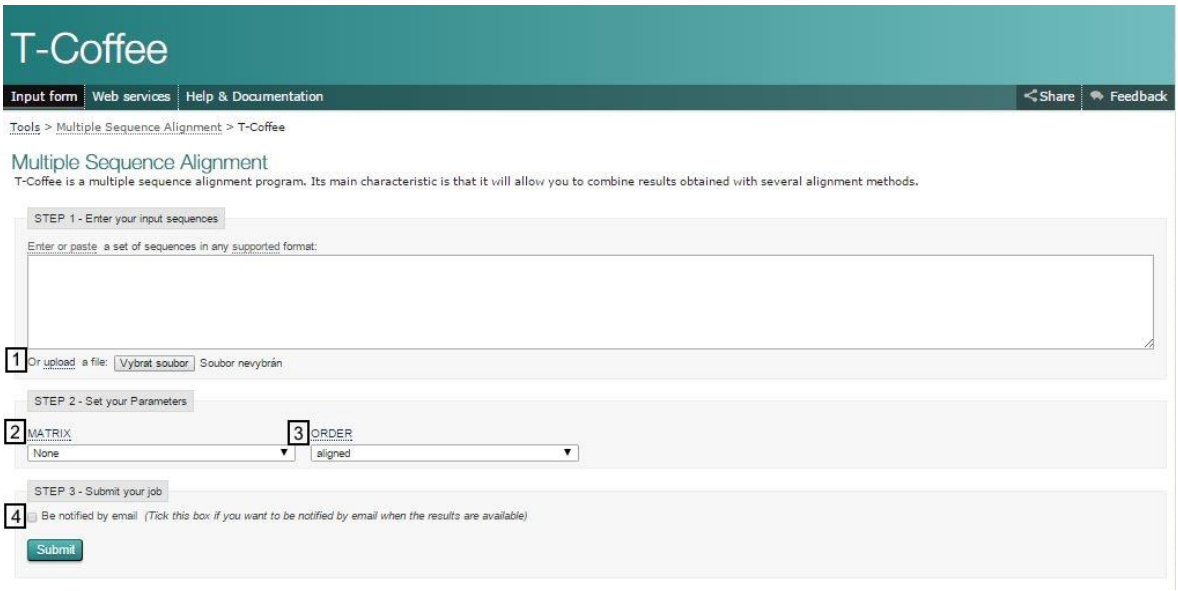
1.2 Nástroj T-Coffee

V roce 2000 byl představen další nástroj pro vícenásobné zarovnání Tree-based Consistency Objective Function for alignment Evaluation se zkratkou T-Coffee. [9]

Zjednodušený algoritmus tohoto nástroje (viz. Příloha II) je následující. Všechny možné dvojice sekvencí jsou zarovnány párovým lokálním i globálním způsobem. Při těchto zarovnáních vznikne mnoho možností, je třeba určit, které zarovnání je nejlepší. Každé dvojici sekvencí je přiřazena váha odpovídající danému zarovnání, vznikne tak schéma vah [10]. V T-Coffee vznikají postupně dvě knihovny. Primární knihovna je sestavena ze sekvencí a jejich vah. V této knihovně už ale nejsou všechna zarovnání. Jsou vybrána pouze ta, jejichž shoda je vyšší než 30 %. [9]

Následným krokem je vytvoření rozšířené knihovny. Je obsáhlejší než primární. Její sestavení je založené na přepočítání vah dvojic sekvencí ve vztahu k dalším sekvencím. Postupně se k zarovnaným sekvencím zarovnává další sekvence, zarovnání získá novou váhu. Knihovna je rozšířena o vztahy více než dvou sekvencí. [9]

V závěrečné fázi je využito informací z rozšířené knihovny a díky progresivnímu zarovnání [11] je vytvořené celkové zarovnání. Postup je následující. Vybrání dvojic, sestavení matice vzdáleností z nich je vytvořen vodící strom pomocí NJ metody [5]. Z vodícího stromu jsou postupně podle seskupení zarovnávány sekvence, přičemž první dvojice je zarovnána běžným způsobem a následující již využívá znalostí z rozšířené knihovny. První dvojice tvoří základ, který již zůstává neměnný a další sekvence jsou k němu přiřazovány. [9]



Obr. 2: Čelní panel online nástroje T-Coffee, výchozí nastavení

Při práci s nástrojem T-Coffee je možné měnit následující parametry a tím ovlivnit výsledek zarovnání. Čísla viz. Obr. 2.

1. Přijatelné vstupní formáty sekvencí: GCG, FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP nebo UniProtKB/Swiss-Prot, omezení velikosti: 500 sekvencí a 1 MB.
2. Možnost pro použití substituční matice pro proteinové sekvence- PAM350 či BLOSUM. Při zarovnávání použití matice může ovlivnit, zda bude vložena mezera, či ne. Matice nám zajistí lepší biologické vlastnosti zarovnání.
3. Možnosti: seřadit (aligned) / vstup (input). Pro možnost seřadit je ve výstupním zarovnání řazení sekvencí pozměněno. Pro možnost vstup je ponecháno pořadí sekvencí v zarovnání shodné, jako pořadí vstupních sekvencí.
4. V případě zájmu, je možné nechat si odeslat výsledky na email.

1.3 Nástroj Mafft

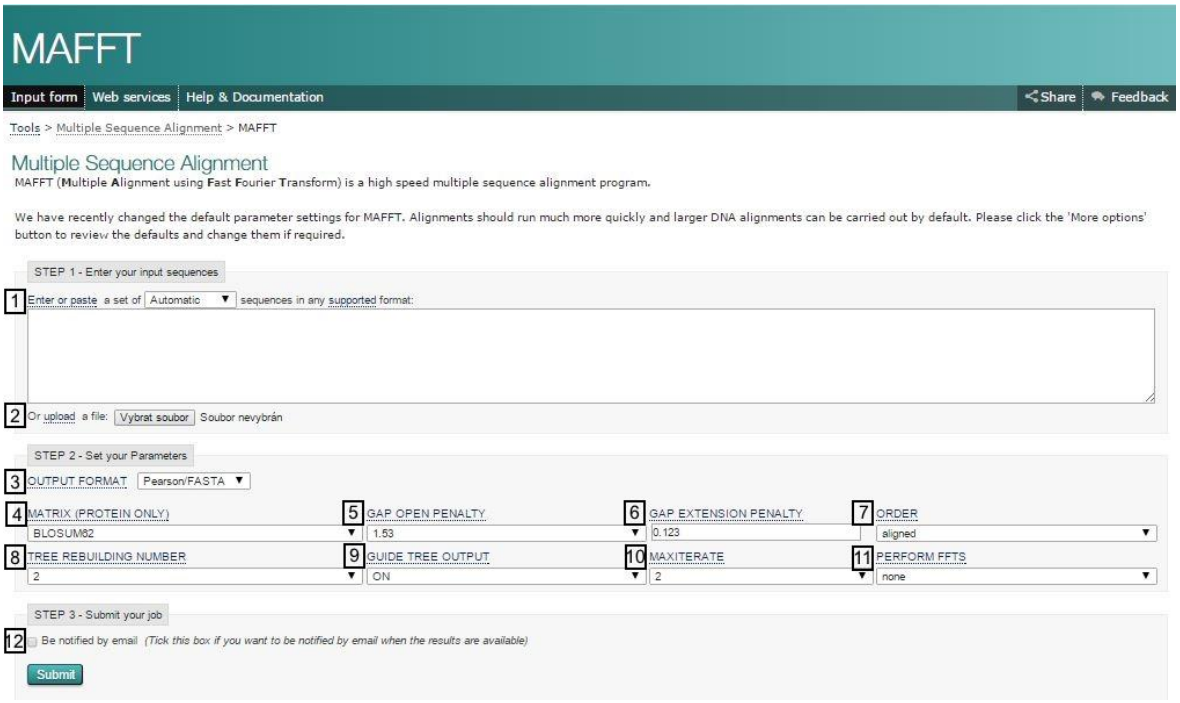
Nástroj pro vícenásobné zarovnání Mafft (Multiple alignment using fast Fourier transform) byl uveden v roce 2002. Jak je patrné z názvu, jeho odlišnost je v použití Fourierovy transformace (FFT) [12]. Mafft se postupně vyvíjel a měl několik verzí, ta nejnovější je Mafft 7. V průběhu let byl zdokonalován, má různé módy a přídatné nástroje pro pozdější úpravu sekvencí (přidání sekvence k zarovnání, odebrání fragmentů, umožňuje omezení zarovnání, paralelní zarovnání,…). [13]

Základem Mafft je již od první verze rychlá Fourierova transformace pro vytvoření skupin podle jejich fyzikálně chemických vlastností. [12]

Homologní sekvence jsou při použití Mafft hledány kombinací FFT (určí, jak jsou sekvence posunuté) a korelace na základě množství a polarity aminokyselin (identifikace

homologních úseků). Dále je vytvořena homologní matice (jak moc si jsou navzájem sekvence podobné) a skórovací systém. Výchozí skórovací matice je 200PAM. Díky těmto upraveným skórovacím systémům je vytvořena matice distancí, z ní pomocí metody UPGMA sestaven vodící strom. Podle větví vodícího stromu jsou sekvence progresivně zarovnány. Algoritmus je zobrazen v Příloze III. [12]

Nejnovější Mafft 7 má hodně nástrojů pro různé typy zarovnání. Nástroje pro zarovnání dlouhých sekvencí jsou NW-NS-PartTree1, NW-NS-PartTree2, NW-NS-DPPartTree1, NW-NS-DPPartTree2. Jsou to progresivní metody, kde místo FFT je použit Needelman-Wunschův algoritmus (NW) [14]. Nástroje pro středně dlouhé zarovnání jsou progresivní metody FFT-NS-1(asi dvakrát rychlejší) a FFT-NS-2 (standardní). Pro krátké sekvence se užívají iterativní metody FFT-NS-i, G-INS-i, L-INS-i, E-INS-i. Poslední automatický nástroj si můžeme zvolit, pokud nevíme, kterou metodu z výše jmenovaných využít. Přiřadí podle délky vstupních dat jednu z metod FFT-NS-2, FFT-NS-i či L-INS-i. [13]



Obr. 3: Čelní panel online nástroje Mafft, výchozí nastavení.

Při práci s nástrojem Mafft je možné měnit následující parametry a tím ovlivnit výsledek zarovnání. Čísla viz. Obr. 3.

1. Možnosti protein/nukleová kyselina. Mafft je schopen automaticky rozeznat typ sekvencí.
2. Vstupní sekvence je možné vkládat ve formátech GCG, FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP nebo UniProtKB/Swiss-Prot, omezení velikosti: 500 sekvencí a 1 MB.

3. Výstupní data je možné požadovat ve formátech: Pearson/FASTA (přednastaveno) nebo ClustalW.
4. Možnost pro použití substituční matice pro proteinové sekvence: BLOSUM30/45/62/80 nebo PAM 100/200. Při zarovnávání použití matice může ovlivnit, zda bude vložena mezera, či ne. Matice nám zajistí lepší biologické vlastnosti zarovnání. Pro BLOSUM platí, že čím menší podobnost předpokládáme, tím nižší číslo matice používáme a u PAM je to naopak.
5. Penalizace zavedení mezery (z angl. gap opening penalty) lze nastavit v rozmezí 1-3. Původně je nastavená hodnota 1,53. Nižší hodnota předpokládá častější vkládání mezer než vyšší hodnota.
6. Penalizace prodlužování mezery (z angl. gap extension penalty) lze nastavit na hodnotu 0-1. Původní hodnota 0,123. Při hodnotě 1 budou mezery prodlužovány méně než při hodnotě 0. S vyšší hodnotou je pravděpodobné, že získáme celkově kratší zarovnání, než s nižší hodnotou.
7. Možnosti: seřadit (aligned) / vstup (input). Pro možnost seřadit je ve výstupním zarovnání řazení sekvencí pozměněno. Pro možnost vstup je ponecháno pořadí sekvencí v zarovnání shodné, jako pořadí vstupních sekvencí
8. Mafft používá přerovnání vodícího stromu, to je možné zakázat, pokud v tomto poli nastavíme, místo předvolené hodnoty 2, hodnotu 1. Zarovnání proběhne zhruba dvakrát rychleji, ale na úkor přesnosti. Lze zde nastavit hodnoty 0 (vodící strom se vůbec nevytvoří), 1,2,5,10,20,50,80 a 100. Je třeba ale pamatovat, že čím víckrát dáme přerovnat strom, tím déle bude zarovnání trvat a ne vždy je to přínosné (zejména u malého množství sekvencí).
9. Možnosti zapnout (on) /vypnout (off). Pokud nechceme, nemusíme paměťový prostor zatěžovat pamatováním si vodícího stromu. Zrychlí se tím algoritmus.
10. Počet zpřesňujících iterací lze nastavit na hodnoty 0,1,2,5,10,20,50,80 a 100. Výchozí nastavení je 0. Při velkém počtu iterací je velice dlouhý výpočetní čas.
11. Tento parametr výrazně ovlivňuje algoritmus. Při změně na lokální (z angl. localpair) je párové zarovnání sekvencí provedeno lokálně pomocí Smith-Watermanova algoritmu. Je pomalejší než výchozí nastavení, ale přesnější. Možnost afinitní zarovnání (podle generalized affine gap cost) je opět pomalejší a přesnější, doporučuje se pro zarovnání do 200 sekvencí s předpokladem velkých vnitřních mezer. Poslední možná volba globální (z angl. globalpair) zarovná sekvence globálním způsobem podle Needlemanova-Wunschova algoritmu. Doporučen je opět do 200 sekvencí s předpokladem globální podobnosti. Opět přesnější ale pomalejší než výchozí nastavení.
12. V případě zájmu, je možné nechat si odeslat výsledky na email.

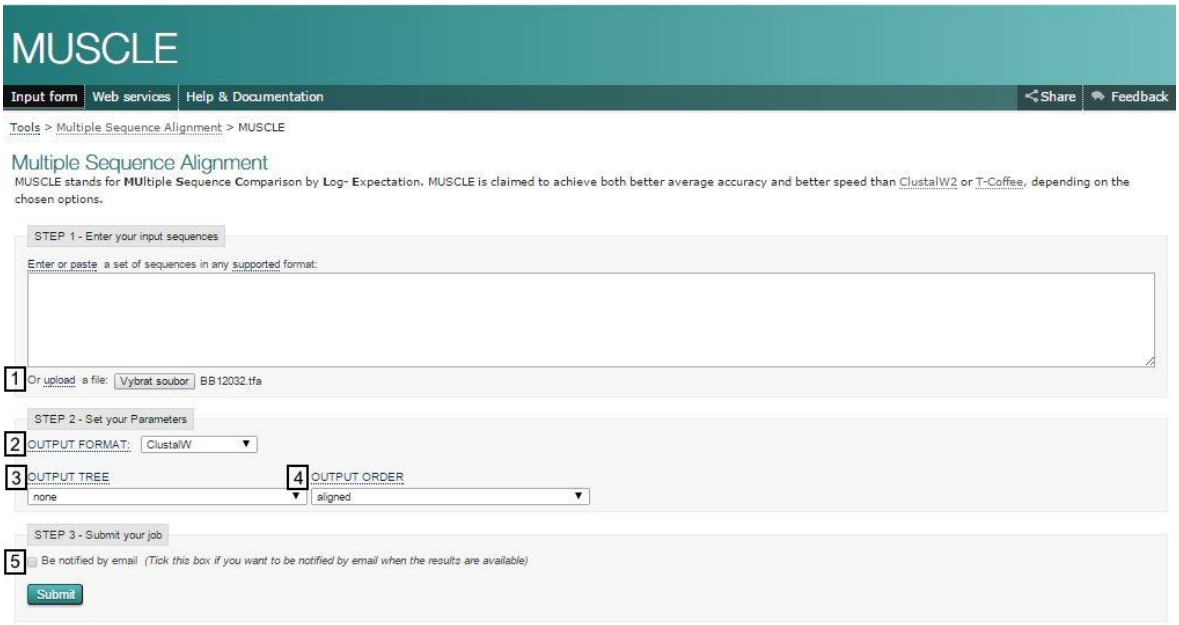
1.4 Nástroj Muscle

Dalším nástrojem pro vícenásobné zarovnání je Muscle (multiple sequence comparison by log expectation). Byl uveden v roce 2004 jako nejrychlejší nástroj pro zarovnání. Má opět více variant pro různé aplikace (Muscle -fast, Muscle -prog a Muscle). [15]

Celkový algoritmus se skládá ze tří stupňů a je zobrazen v Příloze IV. Výhodou je, že po dokončení každého stupně máme k dispozici celkové zarovnání. První stupeň, nazván jako progresivní návrh, se skládá z následujících úkonů: porovnání všech párů sekvencí pomocí k-mer počtů (popsané v [32]) nebo je vytvořeno globální zarovnání páru a je jim určená hodnota částečné identity. Z těchto údajů je sestavena matice distančních vzdáleností a pomocí metody UPGMA nebo NJ algoritmu je sestaven kořenový vodící strom. Pomocí progresivního algoritmu dojde k zarovnání všech sekvencí od kořene ke všem větvím stromu. První stupeň je ukončen a zarovnání je k dispozici. [16]

Druhý stupeň, zlepšující progresivní návrh, vychází z právě vytvořeného zarovnání. Všechny páry v zarovnání jsou ohodnoceny podle podobnosti. Pokud máme tyto hodnoty je sestavena Kimurova matice vzdáleností a dojde k sestavení nového vodícího stromu metodou UPGMA. Nyní máme stromy dva, z prvního a druhého stupně, je nutné jejich porovnání s ohledem na vnitřní uzly v místech, kde se větve vodících stromů liší. Po zjištění odlišností je stávající zarovnání z prvního stupně nejprve zdvojeno a následně jsou upravena místa, kde se stromy liší, jde se od konců větví ke kořeni. [15][16]

Třetí stupeň, „uhlazení“, začíná odstraněním okraje z vodícího stromu z druhého stupně a jeho rozdělení na dvě části. (Odstraňuje se postupně od nejvzdálenějšího místa od kořene.) Pro takto vzniklé části stromů je vypočteno zarovnání, profil. Dalším krokem je zarovnání těchto dvou profilů, čímž vznikne nové, již třetí zarovnání. Pro ohodnocení je vypočítáno skóre sumy párů (SP skóre). Když je SP skóre horší, zarovnání je vymazáno a třetí stupeň se opakuje. Pokud je skóre lepší, zarovnání zůstává ponecháno a celý třetí stupeň probíhá od začátku. K ukončení procesu dochází ve chvíli, kdy SP skóre konverguje, nebo pokud proběhlo určité množství iterací. V případě odstranění všech kořenů je zarovnávání také ukončeno. [15][16]



Obr. 4: Čelní panel online nástroje Muscle, výchozí nastavení.

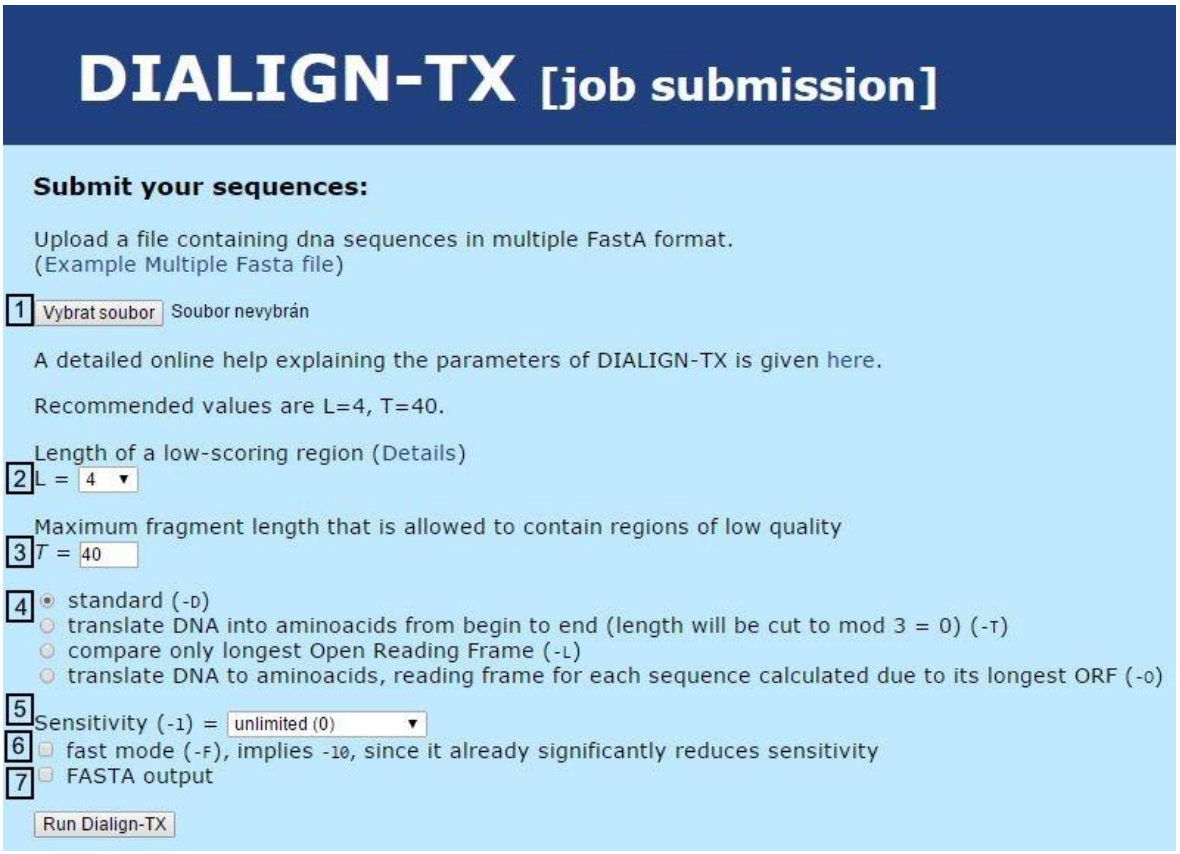
Při práci s nástrojem Muscle je možné měnit následující parametry a tím ovlivnit výsledek zarovnání. Čísla viz. Obr. 4.

1. Vstupní sekvence je možné vkládat ve formátech: GCG, FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP nebo UniProtKB/Swiss-Prot
2. Výstupní data je možné požadovat ve formátech: Pearson/FASTA, ClustalW, ClustalW(strict), HTML, GCG MSF, Phylip interleaved, Phylip sequential.
3. Pokud vyžadujeme u výstupu mít k dispozici vodící stom, je třeba tuto možnost zvolit. Je možné vybrat si vodící strom z první či druhé iterace.
4. Možnosti: seřadit (aligned) / vstup (input). Pro možnost seřadit je ve výstupním zarovnání řazení sekvencí pozměněno. Pro možnost vstup je ponecháno pořadí sekvencí v zarovnání shodné, jako pořadí vstupních sekvencí
5. V případě zájmu, je možné nechat si odeslat výsledky na email.

1.5 Nástroj Dialign

Nástroj Dialign byl uveden v roce 1998, jak pro párové, tak pro vícenásobné zarovnání. Hlavním rozdílem od ostatních zarovnání je, že Dialign nezavádí penalizaci mezer. To je výhodné, zejména pokud si sekvence nejsou globálně podobné. Dialign je dostupný v několika verzích: Anchored Dialign (vícenásobné zarovnání, kde uživatel může zadat vlastní omezení), CHAOS-Dialign (párové i vícenásobné zarovnání, možnost zobrazení pomocí nástroje ABC), Dialign-TX (nahrazuje Dialign-T, „hladký“ algoritmus a progresivní přístup k vícenásobnému zarovnání) a Dialign-PFAM (pro proteinové sekvence, užití databáze PFAM a HMM). [17]

Každá z uvedených verzí se samozřejmě trochu liší, základ zůstává však stejný. Následující popis se vztahuje k Dialign-TX. Ze vstupních sekvencí jsou vytvořeny fragmenty (párová zarovnání všech různých kombinací). Každý fragment má určitou váhovou hodnotu, sestavenou podle toho, jak moc pravděpodobné je vytvoření právě takového fragmentu. Váhová hodnota je také ovlivněna ostatními sekvencemi. Zaznamenává, že motivy v zarovnání dvou sekvencí se vyskytují i v jiných sekvencích. Jako základ zarovnání jsou vybrány ty sekvence, které mají nejvyšší skóre, vyšší pravděpodobnost sestavení. Postupně jsou přidávány další sekvence. V posledním kroku jsou do sekvencí přiřazeny také mezery. Algoritmus zarovnané znaky zapisuje velkými a nezarovnané malými písmeny. Dále v práci je Dialign-tx označován jako Dialign. [18]



Obr. 5: Čelní panel nástroje Dialign-TX, výchozí nastavení

Při práci s nástrojem Dialign je možné měnit následující parametry a tím ovlivnit výsledek zarovnání. Čísla viz. Obr. 5.

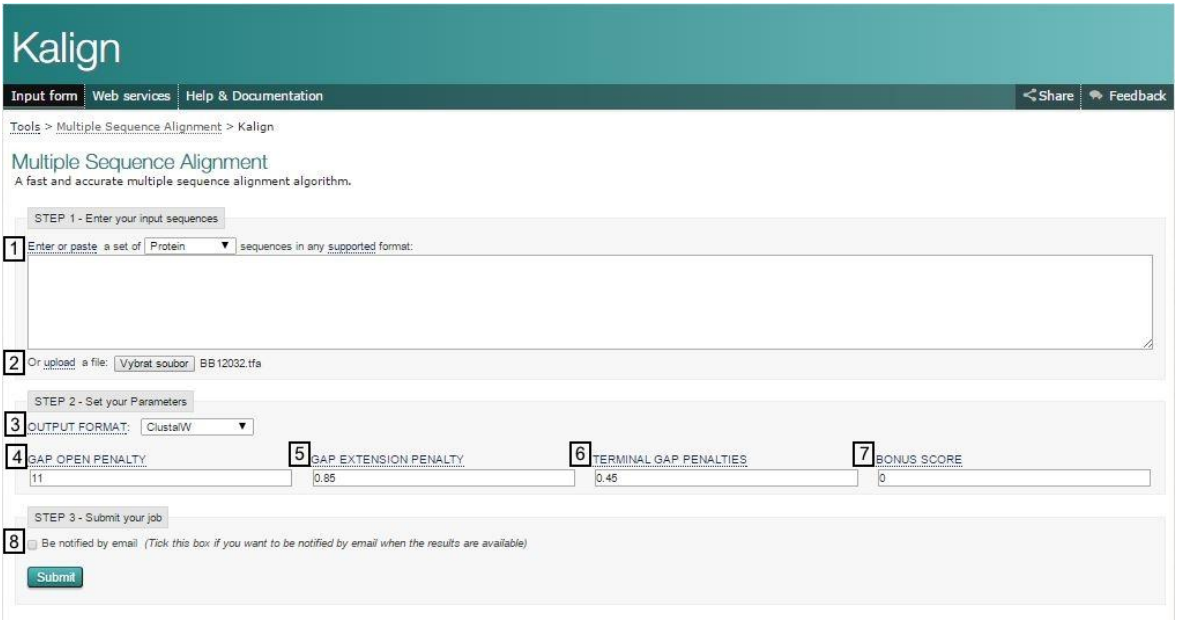
1. Soubory je třeba vkládat ve formátu Multiple Fasta.
2. Parametr značený jako L označuje délku oblasti nízké kvality (za nízkou kvalitu je považována oblast se záporným Needelman-Wunschovým skóre). Výchozí nastavení: L = 4. Jiné možnosti nastavení jsou 0-10. Autor ale doporučuje ponechat pečlivě vybranou hodnotu.
3. Písmenem T se označuje délka fragmentu, ve kterém může být region s nízkou kvalitou zarovnání. Hodnotu je opět možné změnit, ale není to doporučováno.

4. Pro DNA lze použít zarovnání po translaci či porovnání pouze nejdelšího čtecího rámce, případně obojí zároveň.
5. Je možné mít požadavky na sensitivitu programu. Výchozí hodnota je 0 (unlimited), další možnosti jsou 1 (reduced) a 2 (strongly reduced). Citlivost je snížena v prvním kroku párového zarovnávání (1) nebo po celou dobu párového zarovnání (2). Toto snížení nemá vliv na rychlost zarovnání.
6. Pro zkrácení doby zarovnání byl vytvořen tzv. rychlý režim. Je doporučen pro velké množství dlouhých vstupních sekvencí, které nejsou lokálně příbuzné. U tohoto módu je nízká citlivost.
7. Výstupním formátem je textový soubor, v případě zájmu je možné zvolit formát FASTA.

1.6 Nástroj Kalign

Veřejně dostupný nástroje Kalign byl vytvořen v roce 2005. Autoři udávají, že je to rychlý a přesný nástroj s vysokou výkonností pro zarovnání proteinových a nukleotidových sekvencí. [19] [20]

Kalign má analogický přístup k zarovnání, jako progresivní metody. Všechny sekvence jsou zarovnány do páru, je vypočítána matice vzdáleností jednotlivých sekvencí pomocí Muth and Manber algoritmu (původní verze Kalign využívala Wu-Manber algoritmu [19]). Pomocí dot-plot jsou nalezeny podobné sekvence a je vybrána diagonála, od které se dále odvíjí další kroky zarovnání. Následně je sestrojen vodící strom pomocí metody UPGMA, začíná se tvořit nejlépe hodnocenou diagonálou až k těm nejhůře hodnoceným. Z tohoto stromu je již pomocí dynamického programování a za pomoci skórovacího systému sestaveno výsledné zarovnání. [19][20]



Obr. 6: Čelní panel nástroje Kalign, výchozí nastavení.

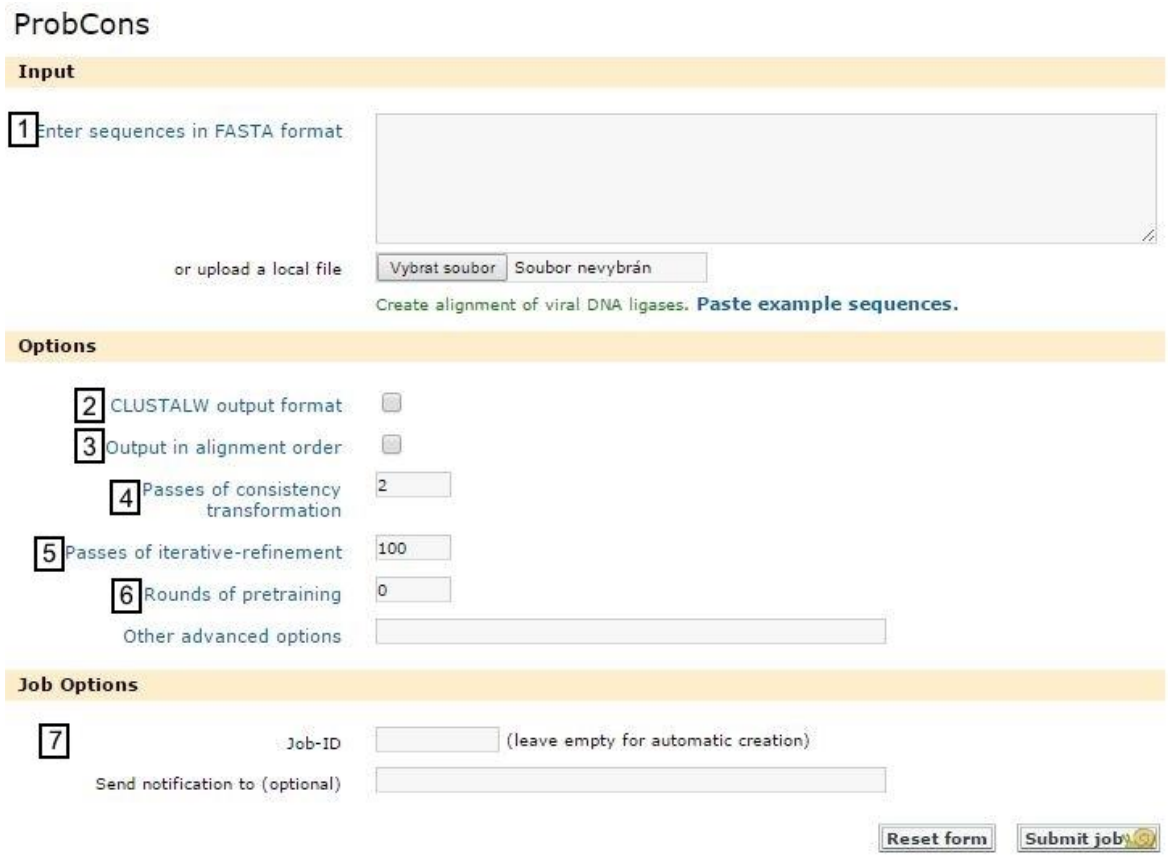
Při práci s nástrojem Kalign je možné měnit následující parametry a tím ovlivnit výsledek zarovnání. Čísla viz. Obr. 5.

1. Možnosti protein/nukleová kyselina volíme podle typu sekvence, jež chceme zarovnat.
2. Vstupní sekvence je možné vkládat ve formátech: GCG, FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP nebo UniProtKB/Swiss-Prot, omezení velikosti: maximální 2000 sekvencí o 2 MB.
3. Výstupní data je možné požadovat ve formátech: Pearson/FASTA, ClustalW (přednastaven) či MACSIM.
4. Penalizace první vložené mezery, gap open penalty, lze nastavit v rozmezí 5-30. Původně je nastavená hodnota 11. Nižší hodnota předpokládá častější vkládání mezer než vyšší hodnota.
5. Penalizace prodlužování mezery, gap extension penalty, lze nastavit na hodnotu 0,2-5. Původní hodnota 0,85. Při hodnotě 5 budou mezery prodlužovány méně než při hodnotě 0,2. S vyšší hodnotou je pravděpodobné, že získáme celkově kratší zarovnání, než s nižší hodnotou.
6. Penalizace poslední vložené mezery, terminal gap penalty, lze nastavit v rozmezí 0-5. Původně je nastavená hodnota 0,45. Nižší hodnota předpokládá častější ukončování mezer než vyšší hodnota.
7. Bonusové skóre (v rozmezí 0-5,2 pro proteiny a -125-0 pro nukleotidy) je konstanta, která je přičtena ke každému prvku substituční matice. Tento prvek má zvýhodnit nukleotidy či aminokyseliny před mezerami v zarovnání. Je doporučeno tuto hodnotu měnit, pokud si nejsme jisti homologií.
8. V případě zájmu, je možné nechat si odeslat výsledky na email.

1.7 Nástroj ProbCons

ProbCons (Probabilistic consistency-based multiple semence alignment) další z veřejně dostupných nástrojů pro vícenásobné zarovnání proteinů, vytvořený v roce 2005, se vyznačuje zejména přihlížení k pravděpodobnosti složení zarovnání proteinů. [21]

Algoritmus ProbCons se dost liší od ostatních, běžně užívaných, nástrojů pro vícenásobné zarovnání. K zarovnání sekvencí do páru je využito HMM. Vzniklé páry mají přiděleny hodnoty očekávané přesnosti, z nichž je vytvořen shlukováním vodící strom. Další vypočítanou hodnotou, ještě z párového zarovnání je skóre kvality shody. Progresivní zarovnání probíhá na základě vodícího stromu s přihlédnutím ke skóre kvality shody. Zarovnání je hodnoceno pomocí metody sumy párů, přičemž penalty mezer jsou nastaveny na nulovou hodnotu. Konečné zarovnání je to, které je nejlépe ohodnoceno. [21]



Obr. 7: Čelní panel online nástroje ProbCons, výchozí nastavení.

Při práci s nástrojem ProbCons je možné měnit následující parametry a tím ovlivnit výsledek zarovnání. Čísla viz. Obr. 7.Obr. 5

1. Vstupní formát musí být FASTA.
2. Je možné vyžadovat výstupní zarovnání ve formátu CLUSTALW, jinak je výstupem soubor MFA.
3. Po označení této možnosti je pořadí v zarovnání shodné s pořadím vložených sekvencí.

4. Při tvorbě zarovnání je možné zvolit, kolikrát bude skóre přepočítáno. Výchozí hodnota je 2 (jiné možnosti 0,1,3)
5. Počet, kolikrát bude ověřeno zarovnání rozdělením do dvou skupin a následným přerovnáním. Výchozí hodnota je 0 (možné je 0 až 1000).
6. Kola přetrénování před zarovnáním je možné využít v případě, že nám nevyhovují pro dané sekvence standardní parametry. Jindy se toto ověření nedoporučuje, může způsobit nestabilitu zarovnání. Původně je nastavená 0 až 20.
7. V případě zájmu, je možné nechat si odeslat výsledky na email.

1.8 Funkce multialign

Součástí bioinformatického toolboxu programu Matlab je funkce *mulialign*. Díky této funkci je možné vícenásobně zarovnat sekvence pomocí progresivního zarovnání. Ohodnocení párových zarovnání ignoruje mezery a shody či neshody znaků jsou ohodnoceny substituční maticí, výchozí nastavení je Gonnet pro aminokyseliny a NUC 44 pro nukleotidy. Zarovnání probíhá párově hierarchicky podle vodícího stromu, který je tvořen pomocí NJ algoritmu. Nastavení funkce je možné přizpůsobit vlastním potřebám velkým množstvím parametrů, jako jsou substituční matice, metoda sestavení vodícího stromu, penalizace zavedení a prodloužení mezery, a další. [37]

2 Parametry pro popis kvality vícenásobného zarovnání

Vícenásobné zarovnání patří mezi základní nástroje pro práci se sekvencemi v molekulární biologii. Je to zásadní způsob jak pochopit souvislosti mezi strukturou a funkcí proteinových rodin. Samotná predikce struktury by se bez správného zarovnání také neobešla. U porovnání více než dvou sekvencí znatelně stoupá s každou přidanou sekvencí náročnost výpočtu. Je obdivuhodné, že je možné zarovnání jako takové provést, a že nástrojů pro vícenásobné zarovnání je poměrně hodně. Díky tomu si můžeme vybírat vhodný nástroj, podle toho, co je pro nás v daném zarovnání prioritou. Tato kapitola nabízí ucelený přehled parametrů pro popis kvality vícenásobného zarovnání.

2.1 Celkové skóre zarovnání

Jako první parametr pro zarovnání uvádím hodnocení zarovnání jako celku. V tomto přístupu jsou brány v potaz zejména shody či neshody prvků.

Jednou z možností takového typu je metoda sumy párů (SP, z anglického výrazu sum of pair score). V tomto případě jsou jednotlivé prvky zarovnání porovnávány ve sloupci vzájemně mezi sebou. Při shodě párů ve sloupci skóre narůstá. Pokud máme i referenční zarovnání je možné skóre sumy párů vypočítat i pro ni a následně normalizovat skóre pro námi vytvořené zarovnání. Díky normalizaci dosáhneme objektivnějšího hodnocení zarovnání. [21]

$$SP = \sum_{i=1}^{M} S_i \,, \tag{1}$$

kde M je počet sloupců zarovnání a S_i je suma párů pro jeden sloupec, vypočítá se následovně:

$$S_i = \sum_{j=1, j\neq k}^{N} \sum_{k=1}^{N} p_{ijk} \,, \tag{2}$$

kde N je počet sekvencí a p_{ijk} nabývá hodnot podle nastavení skórovacího systému (hodnota shody, neshody, přítomnost mezery, použití substituční matice).

```
Clustal omega

1.sekvence          QWCCSCDNREEPPTAPS--
2.sekvence          RNECSCDNTSEPAPS----
3.sekvence          ILCCSCDNPPYPS------
4.sekvence          --NCSCDNREEPTSP----
5.sekvence          -FCCSCADNGITLHEPPPS
```

Obr. 8: Zarovnání pro ukázky výpočtu

Ukázka výpočtu sumy párů pro zarovnání na Obr. 8

Skórovací systém: substituční matice Blosum62

penalizace mezer p_{ijk} = -2, při přítomnosti dvou mezer p_{ijk} = 0

$$S_1 = p_{1QR} + p_{1QI} + p_{1Q-} + p_{1Q-} + p_{1RI} + p_{1R-} + p_{1R-} + p_{1I-} + p_{1I-} + p_{1--} =$$
$$= 1 - 3 - 2 - 2 - 3 - 2 - 2 - 2 - 2 + 0 = -17$$
$$S_2 = p_{2WN} + p_{2WL} + p_{2W-} + p_{2WF} + p_{2NL} + p_{2N-} + p_{2NF} + p_{2L-} + p_{2LF} + p_{2-F} =$$
$$= -4 - 2 - 2 + 1 - 3 - 2 - 3 - 2 + 0 - 2 = -19$$

Obdobně další sloupce:

$$S_3 = 6;\ S_4 = 90;\ S_5 = 40; S_6 = 90;\ S_7 = 28;\ S_8 = 40;\ \ S_9 = -4;\ \ S_{10} = -4$$
$$S_{11} = -1;\ S_{12} = 38;\ S_{13} = -8;\ S_{14} = -14;\ \ S_{15} = -11; S_{16} = -5;\ \ S_{17} = -13$$
$$S_{18} = -8; S_{19} = -8$$

Výsledné skóre pomocí sumy párů:

$$SP = \sum_{i=1}^{19} S_i = -17 - 19 + 6 + 90 + 40 + 90 + 28 + 40 - 4 - 4 - 1 + 38 - 8 - 14$$
$$- 11 - 5 - 13 - 8 - 8 = 220$$

Sloupcové skóre (CS, z anglického výrazu column score) hodnotí, jak program dokáže zarovnat celý sloupec. Zde se pracuje s celými sloupci. Skóre se navyšuje, pokud je celý sloupec shodný. Opět je zde možná normalizace za předpokladu existence referenčního zarovnání. [21][22]

$$CS = \frac{\sum_{i=1}^{M} C_i}{M}, \tag{3}$$

kde *M* je počet sloupců zarovnání a *Ci* = 1, pokud je znak zastoupen ve všech sekvencích a *Ci* = 0 v případě i jediné odlišnosti.

Ukázka výpočtu sloupcového skóre pro zarovnání na Obr. 8

$$CS = \frac{\sum_{i=1}^{19} C_i}{19}$$
$$CS = \frac{0 + 0 + 0 + 1 + 1 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0}{19}$$
$$CS = \frac{3}{19} \cong 0{,}158$$

2.2 Časová náročnost

V anglické literatuře je uváděná časová náročnost pod pojmem CPU time (central processing unit time). Časová náročnost je téměř synonymem složitosti algoritmu. Čím je algoritmus náročnější, tím déle trvá vytvoření vícenásobného zarovnání. Jednotlivé nástroje náročnost uvádí ve vztahu k počtu sekvencí (N) a jejich délkám (L). Při užití dynamického programování je náročnost algoritmu uváděna jako O ($2^N L^N$). Náročnost algoritmů zarovnání již nemá N v mocnině (např. Muscle je O (N^4+NL^2) [15]). Při porovnání s náročností použití pouze dynamického programování je možno vidět značný postoj kupředu. Jsou vyvíjeny stále výkonnější systémy, i přesto je žádoucí, aby nástroj byl co nejméně náročný, i co se paměti týče. [21]

2.3 Biologické zohlednění

Biologické hledisko se zdá být doplňkové, ale je nutné pamatovat na to, že molekulární biologie je jeden z důvodů proč zarovnání vytváříme.

Při evoluci dochází v genomu k mutacím. Ty se mohou projevit jako inzerce, delece či substituce. Substituce je z evolučního hlediska pravděpodobnější. Je důležité na to brát zřetel a zbytečně do zarovnání nevkládat mezery (o tomto problému níže viz. 2.4). Dále není substituce jako substituce. Změna nukleotidu či aminokyseliny v jinou nemá vždy stejnou pravděpodobnost. Daleko pravděpodobnější substituce je v rámci nukleotidů/aminokyselin podobných vlastností. Tento parametr by měl být řešen v každém nástroji užitím substituční matice (např. Nuc 44, PAM, Blosum, Gonnet, viz. [23], [24] a [25]) při vytváření ohodnocení skóre všech zarovnaných párů. U některých nástrojů je možno si matici zvolit, u jiných je předvolena. Při substituci některých nukleotidů se nemusí změnit struktura ani funkce výsledného proteinu, při jiné substituci může mutace způsobit výraznou odlišnost.

2.4 Použití mezer, zarovnání v bloku

Pokud se zaměříme na menší části zarovnání, můžeme si všímat, jak často a jak dlouhé jsou vkládané mezery, celistvosti/rozkouskovanosti zarovnání, zda jsou v zarovnání výrazné bloky a dalších skutečností.

Mezery, důsledek inzercí či delecí, jak je zmíněno výše, nejsou tak pravděpodobné jako substituce. Proto při zarovnávání bývají skórovacím systémem výrazně znevýhodněny. Zejména vytvoření mezery. Prodloužení již započaté mezery je hodnoceno mírněji. Pro tyto parametry se používají anglické termíny gap opening (zavedení mezery) a gap extension (prodloužení mezery). U mnoha nástrojů si uživatel tyto hodnoty může sám určit.

V závislosti na vkládání mezer mohou být v zarovnání vytvořeny bloky.

Při zarovnávání jsou bloky předmětem našeho zájmu, chceme najít to, co mají sekvence společné, podobné. U bloků se dá hodnotit délka, podobnost či četnost bloků v zarovnání. Celkové hodnocení bloku (BCS z anglického block column score) vyjadřuje téměř to samé jako CS, pouze se vztahuje k bloku. [22]

$$BCS = \frac{\sum_{i=1}^{m_b} C_{bi}}{m_b}, \tag{4}$$

kde M_b je počet sloupců v bloku a $C_{bi} = 1$, pokud jsou všechny znaky ve sloupci stejné a $C_{bi} = 0$ v případě i jediné odlišnosti.

2.5 Entropie

Entropie může vyjadřovat množství informace vztažené ke slovu, pojem také označuje míru neuspořádanosti systému. Početně se dá určit pomocí následujícího vzorce:

$$H_i = -\sum_a f_{a,i} * \log_2 f_{a,i}\,, \tag{5}$$

kde *Hi* je entropie jedné pozice zarovnání (jedné pozice konsenzuálního řetězce, respektive entropie celého sloupce), a jsou všechny možnosti výskytu na pozici (pro DNA a=A,C,G,T,-; pro RNA a=A,C,G,U,-; pro protein a = A,R,N,D,C,Q,E,G,H,I,L,K,M,F, P,S,T,W,Y,V,B,Z,X,-), $f_{a,i}$ je frekvence výskytu prvku a na pozici *i*, lze vypočítat pomocí vzorce $f_{a,i} = n_a/n$ (n_a je počet daného prvku ve sloupci a n je celkový počet znaků ve sloupci, tj. počet sekvencí). Celková entropie zarovnání lze určit průměrem entropií všech sloupců. [26]

Ukázka výpočtu entropie pro některé sloupce zarovnání z Obr. 8

První sloupec obsahuje Q, R, I a dvě mezery.

$$H_1 = -(f_{Q,1} * \log_2 f_{Q,1} + f_{R,1} * \log_2 f_{R,1} + f_{I,1} * \log_2 f_{I,1} + f_{-,1} * \log_2 f_{-,1})$$

$$H_1 = -(\tfrac{1}{5} * \log_2 \tfrac{1}{5} + \tfrac{1}{5} * \log_2 \tfrac{1}{5} + \tfrac{1}{5} * \log_2 \tfrac{1}{5} + \tfrac{2}{5} * \log_2 \tfrac{2}{5}) = 1{,}9219 \text{ bit}$$

Druhý sloupec obsahuje W, N, L, mezeru a F

$$H_2 = -(f_{W,2} * \log_2 f_{W,2} + f_{N,2} * \log_2 f_{N,2} + f_{L,2} * \log_2 f_{L,2} + f_{-,2} * \log_2 f_{-,2} + f_{F,2} * \log_2 f_{F,2})$$

$$H_2 = -\left(\frac{1}{5} * \log_2 \frac{1}{5} + \frac{1}{5} * \log_2 \frac{1}{5} + \frac{1}{5} * \log_2 \frac{1}{5} + \frac{1}{5} * \log_2 \frac{1}{5} + \frac{1}{5} * \log_2 \frac{1}{5}\right)$$

$$H_2 = 2{,}3219 \text{ bit}$$

2.6 Z-skóre

Z-skóre je statistický nástroj, který můžeme použít pro hodnocení párového zarovnání. Při vytvoření průměru z jednotlivých párů zarovnání získáme celkové skóre vícenásobného zarovnání. Z-skóre pro pár sekvencí získáme použitím následujícího vzorce:

$$Z = \frac{S - \bar{x}}{\sigma(x)}, \tag{6}$$

kde S je skóre podle Smith-Watermana (popsáno v [31]), $\bar{x}$ je průměr S skóre a $\sigma(x)$ je směrodatná odchylka. Z-skóre je normalizovaná hodnota průměrů Smith-Watermanova skóre [30]. Ukázka výpočtu je uvedena dále v práci.

3 Metodika testování

Hodnocení zarovnání jednotlivých nástrojů je možné provést s vhodně sestaveným algoritmem, s použitím různých setů sekvencí pro testování s různými vlastnostmi a v dostatečném množství, aby výsledky bylo možné považovat za důvěryhodné. Důležité je také dobře pojmout a vyhodnotit data z algoritmu vzešlá.

3.1 Data pro testování

Pro testování programu i nástrojů byly použity datasety z databáze BaliBase a data z porovnávacího RNA webu (z anglického The Comparative RNA Web, zkratka CRW), který obsahuje mimo jiné data zarovnaných rRNA sekvencí, informace o strukturách sekvencí, o frekvenci a konzervovanosti nukleotidů. [34]

BaliBase

Proteinové sekvence pro testování nástrojů jsou staženy z databáze BaliBase 3 (dostupné z: http://www.lbgi.fr/balibase). Ručně zkontrolovaná zarovnání jsou dostupné v několika formátech (.msf, .rsf, .xml), nezarovnané sekvence jsou přiložené ve formátu fasta (.tfa). BaliBase 3 obsahuje 5 referenčních sad zarovnání. V Tab. 1 jsou uvedeny specifikace těchto skupin. Databáze obsahuje pouze aminokyselinové sekvence. [28]

Tab. 1: Použité sekvence pro testování nástrojů

Označení skupiny	Počet zarovnání	Podnázev	Charakteristika sady
RV11	38	Délka a odlišnost	Sekvence odlišných délek a podobností (shoda < 20 %).
RV12	44		Sekvence odlišných délek a podobností (shoda 20-40 %).
RV20	41	„sirotci“	Sekvence z proteinové rodiny a jedna (či více) vzdálenější sekvence.
RV30	30	„podrodiny“	Sekvence z několika rodin (shoda cca 25 %).
RV40	49	prodloužení	Sekvence s velkými přesahy.
RV50	16	inzerce	Sekvence obsahující inzerce (dlouhé nezarovnatelné úseky).

Data z porovnávacího RNA webu

Nukleotidové sekvence byly vybrány z databáze proteinů vytvořenou Univerzitou Texasu v Autstinu, CRW za účelem porovnávání prostorového uspořádání.

Databáze obsahuje RNA sekvence různých organismů, typů, velikostí, které jsou ručně zarovnané, lze stáhnout tytéž sekvence (nezarovnané). Dále jsou v setech obsaženy tzv. děliče (v originále divider), které naznačují právě prostorové uspořádání. Ty byly pro potřeby této práce nadbytečné. Před dalším zpracováním byly odstraněny. [34] Nukleotidové sekvence jsou obecně delší než proteinové, proto byly vybrány sady dat,

které neměly nadměrnou velikost (horní mez 1 MB) a maximálně 500 sekvencí k zarovnání.

V Tab. 2 jsou uvedeny použité sekvence pro testování vybraných nástrojů vícenásobného zarovnání. Všechny jsou ze sekce primárně zarovnaných, která je podle autorů jejich nejlépe zarovnaná sekce. [34]

Tab. 2: RNA sekvence použité pro testování nástrojů z CRW

Druh zarovnání	Název zarovnání	Druh rRNA	Počet sekvencí	Přibližná délka sekvencí
Archea	5S.A.fasta	5s	147	130bp
Chloroplast	5S.C.fasta	5s	199	120bp
Mitochondrie	5S.M.fasta	5s	49	120bp

Použité nástroje

K zarovnání výše uvedených sekvencí bylo použito 7 veřejně dostupných nástrojů (viz. Tab. 3) a funkce multialign v programu Matlab. V Tab. 4 je uvedené nastavení, v jakém byly jednotlivé metody vícenásobného zarovnání použity.

Tab. 3: Použité nástroje a jejich umístění

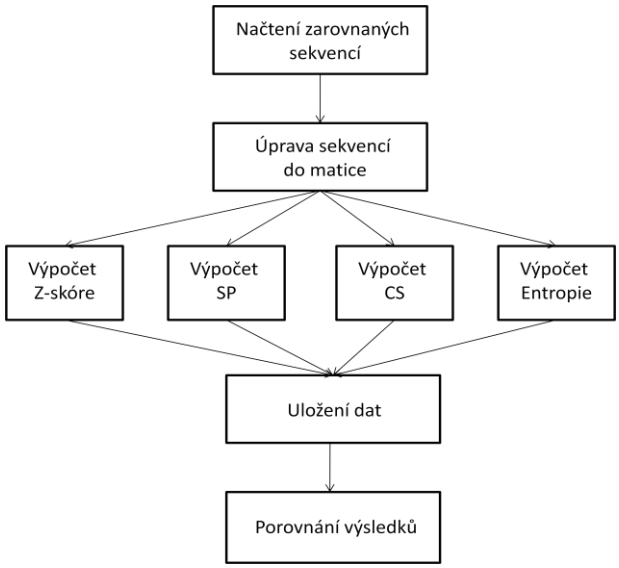
Nástroj	Odkaz
Clustal Omega	http://www.ebi.ac.uk/Tools/msa/clustalo/
T-Coffee	http://www.ebi.ac.uk/Tools/msa/tcoffee/
Mafft	http://www.ebi.ac.uk/Tools/msa/mafft/
Muscle	http://www.ebi.ac.uk/Tools/msa/muscle/
Dialign	http://dialign-tx.gobics.de/
Kalign	http://www.ebi.ac.uk/Tools/msa/kalign/
ProbCons	http://toolkit.tuebingen.mpg.de/probcons

Tab. 4: Výchozí, použité, nastavení veřejně dostupných algoritmů pro vícenásobné zarovnání

Nástroj	Vytvoření vodícího stromu	Zavedení mezery	Prodloužení mezery	Substituční matice
Clustal Omega	Algoritmus k-průměrů/UPGMA	6	1	Gonnet
T-Coffee	NJ metoda	nepoužívá	nepoužívá	nepoužívá
Mafft	UPGMA	1,53	0,123	Blosum 62
Muscle	UPGMA/ NJ metoda	-400 pro DNA; -420 pro RNA; -2,9 pro protein	0	200PAM; Kimurova matice
Dialign	UPGMA	nepoužívá	nepoužívá	Blosum 62
Kalign	Muth a Mamber algoritmus	11	0,85	HOXD,Gonet
Probcons	UPGMA/ NJ metoda	0	0	Blosum 62
Matlab	NJ metoda	nepoužívá	nepoužívá	Gonnet

3.2 Nástroj pro testování kvality vícenásobného zarovnání

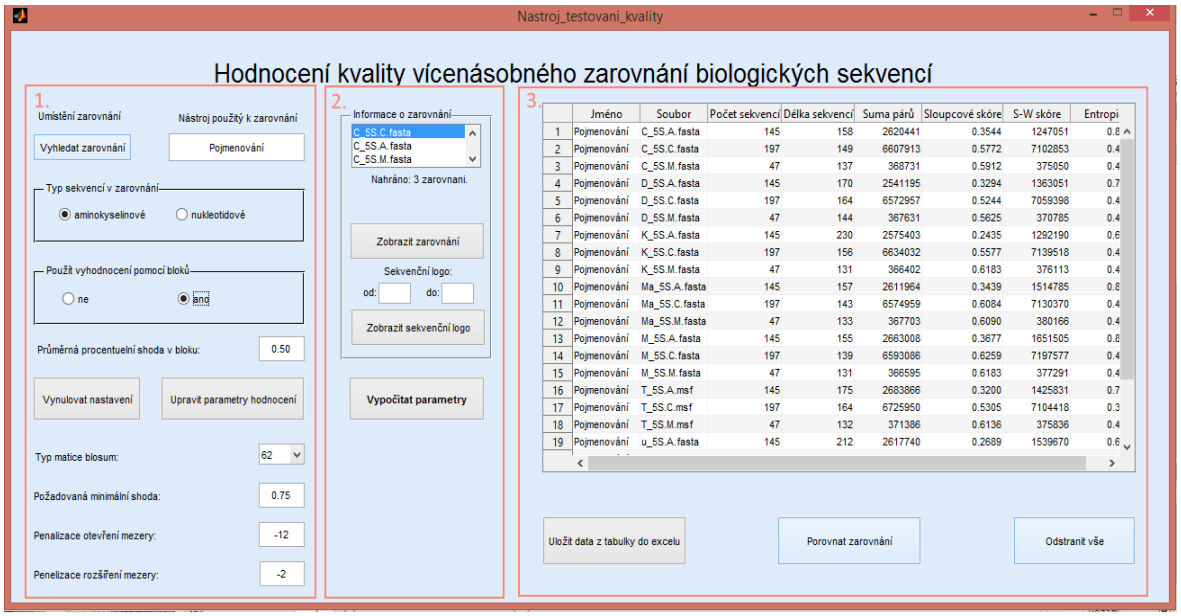
Nástroj pro testování kvality vícenásobného zarovnání je sestaven pro určení nejvhodnějšího vícenásobného zarovnání. Uživatel má různá zarovnání jedné sady sekvencí (např. z různých nástrojů, různých nastavení). Pomocí programu může zjistit, jaké zarovnání by bylo pro jeho použití nejvhodnější. Schéma algoritmu programu je na Obr. 9.



Obr. 9: Blokové schéma algoritmu pro hodnocení kvality vícenásobného zarovnání

Uživatelské rozhraní

Nástroj testování kvality je program určený pro Matlab2012a a vyšší verze. Uživatelské rozhraní naběhne po spuštění skriptu Nastroj_testovani_kvality.m. Na Obr. 10 je hlavní okno programu. První část (označená 1.) se týká načtení zarovnání, prostřední (2.) informací o načteném zarovnání a pravá (3.) ukládání a zobrazení dat a jejich ohodnocení.

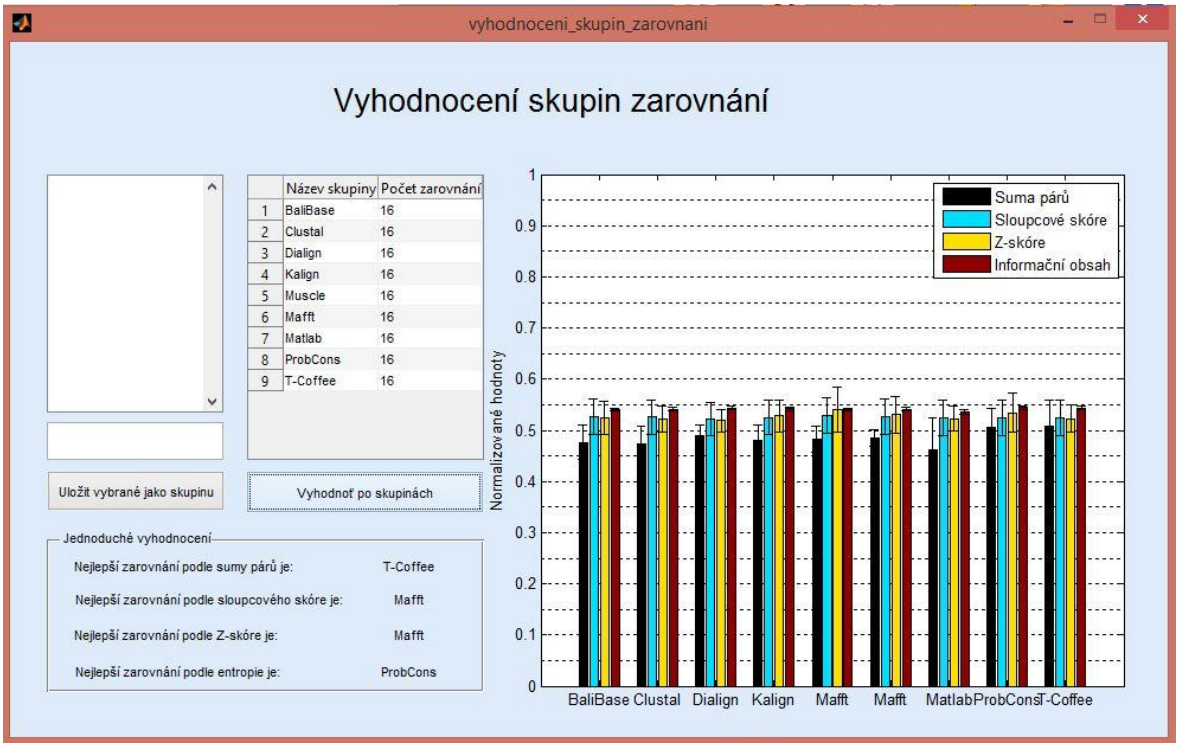


Obr. 10: Uživatelské rozhraní nástroje pro hodnocení vícenásobného zarovnání

Výstupem programu je porovnání zarovnání. Toto porovnání je možné dvěma způsoby. První je vhodný pro jedno zarovnání různými metodami (příp. nastavením, programy,…). Druhý je určen pro hodnocení sady podobných sekvencí, tedy pro určení nejlepšího nástroje, např. pro krátké podobné sekvence. Na Obr. 11 je možné vidět výsledek testování pro druhý typ porovnání. Graf v pravé části okna zarovnání je sestaven pomocí funkce *barwitherr* stažené a dostupné z [36]. Hodnoty výšky sloupců jsou dány průměrnou normalizovanou hodnotou a rozpětí (tzv. errorbars) udává směrodatnou odchylku hodnot. V grafu je místo entropie uváděn informační obsah, aby bylo možné posuzovat všechny parametry způsobem čím vyšší hodnota, tím lepší výsledek. Přepočet entropie na informační obsah (*io*) je proveden pomocí vzorce:

$$io = \frac{\log_2(d) - entropie}{\log_2(d)}, \tag{7}$$

kde *d* je 4 pro nukleotidy a 24 pro aminokyseliny.



Obr. 11: Vyhodnocení testování skupin zarovnání

Podrobnější představení programu lze nalézt v uživatelské příručce uvedené v Příloze VIII.

3.3 Funkce jednotlivých parametrů

Funkce SP.m

Funkce SP.m vypočítá SP skóre pomocí metody sumy párů. Výpočet je proveden stejně, jako je uvedeno v kapitole 2.1. Penalizace jednotlivých případů, které mohou nastat, je dána substituční maticí Blosum (pro proteiny) případně NUC44 (pro nukleotidy). Matice jsou rozšířeny o sloupec zahrnující mezeru. Typ matice Blosum i penalizace mezery je zadána uživatelem. Přednastavené hodnoty programu jsou matice Blosum 62 a penalizace mezer nastavena na hodnotu -14 pro protein a -20 pro nukleotidy.

Funkce CS.m

Funkce CS.m vypočítá sloupcové skóre, jak je uvedeno v kapitole 2.1 s rozdílem hodnocení shody znaku. $C_i = 1$, pokud je zastoupení jednoho znaku větší než uživatelem zadaná procentuální hodnota, jinak je $C_i = 0$. Přednastavená hodnota programu je požadavek větší shody než 75 %.

Ukázka výpočtu sloupcového skóre použitého ve funkci pro zarovnání pro zarovnání na Obr. 8.

$$CS = \frac{\sum_{i=1}^{19} C_i}{19}$$

$$CS = \frac{0+0+0+1+1+1+1+1+0+0+0+1+0+0+0+0+0+0+0}{19}$$

$$CS = \frac{6}{19} \cong 0{,}315789$$

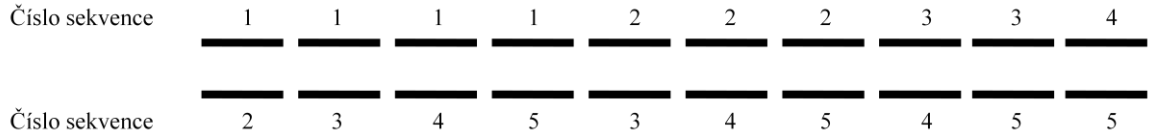
Funkce Entropie.m

Funkce Entropie.m realizuje výpočet entropie tak, jak je uvedeno dříve v kapitole 2.5. Pro stanovení frekvence výskytu jednotlivých prvků (aminokyselin či nukleotidů) byla sestavena funkce *pocetamk.m* a *pocetnukl.m* Frekvence výskytu mezer je realizováno kombinací funkcí *findstr* a *length*.

Funkce Z_score.m

Z_score.m je funkce, která vypočítá z-skóre podle vzorce (6). Důležitý je výpočet Smith-Watermanova skóre (dále SW skóre). To je počítáno pouze pro dvě sekvence, avšak musí být vypočteno pro všechny kombinace dvojic sekvencí. Proto bylo zvoleno sestavení jedné dlouhé sekvence, která obsahuje veškeré možnosti párů sekvencí, příklad pro 5 sekvencí je znázorněn na Obr. 12. Pro takto sestavené zarovnání je jednoduše vypočítáno SW skóre. Pro zarovnání každého z použitých nástrojů získáme jinou hodnotu SW skóre.

Z-skóre pro každé zarovnání je získáno normalizací v rámci těchto hodnot, respektive dosazením do vzorce (6) v kapitole 2.6. Pro názornost je dále uveden příklad výpočtu. Penalizační aparát pro SW skóre v programu nastaven následovně. Pro proteiny je použita matice Blosum 62 (možnost volby), penalizace zavedení a rozšíření mezery na hodnoty -12 a -2, pro nukleotidy je použita matice NUC44 a penalizace mezer -16 a -4. [35]



Obr. 12: Sestavení jedné sekvence všech možných kombinací párů

Ukázka výpočtu Z-skóre v algoritmu

Výpočet SW skóre

- skórovací systém: shoda=2, neshoda=-1, mezera=-3 (SW skóre je vždy 0 nebo kladné)

```
1. nástroj     Q W C C S C D  N R E  E  P  P  T A P S
               R N E C S C D  N T S  E  P  A  P S - -
SW             0 0 0 2 4 6 8 10 9 8 10 12 11 10 9 6 3

1. nástroj     Q W - C C S C D  N  R  E  E  P  P T A P S
               R N E C C S C D  N  P  P  Y  -  - - - P S
SW             0 0 0 2 4 6 8 10 12 11 10 9  6  3 0 0 2 4

1. nástroj     Q W - C C S C D  N  R  E  E P  P  T  A P  S
               R N E C C S C D  N  -  -  - P  P  Y  - P  S
SW             0 0 0 2 4 6 8 10 12 11 10 9 11 13 12 9 11 13
```

Výpočet z-skóre

$\overline{x}$=6,667; μ=4,497

Z_1=(3-6,667)/ 4,497= -0,815

Z_2=(4-6,667)/ 4,497= -0,593

Z_3=(10-6,667)/ 4,497=1,408

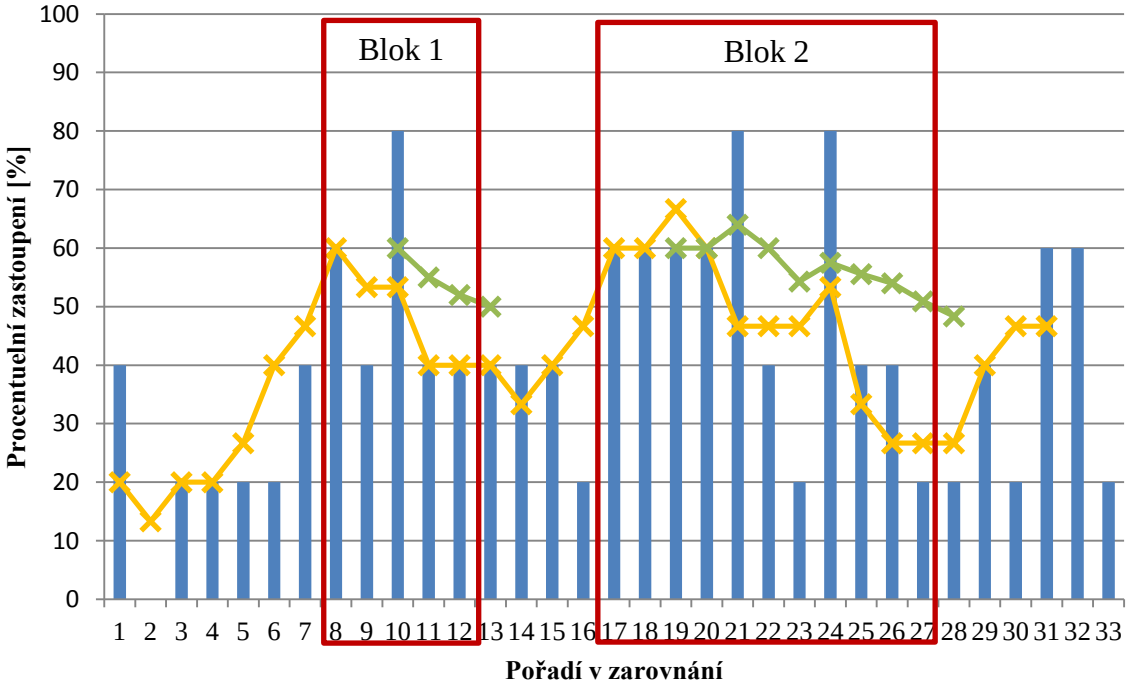
Funkce pro sestavení bloku

Veškeré dříve jmenované parametry lze vypočítat dvěma způsoby. První možností je vypočítat hodnoty z celkového zarovnání, tak jak je. Druhá možnost se snaží zohlednit, jak je zarovnání dobře sestavené v oblastech, kde není třeba doplňovat ve většině sekvencí mezery. Před samotným výpočtem jsou vytvořené bloky, které splňují uživatelem nastavenou průměrnou shodu v bloku.

Jako blok jsou určeny minimálně 3 sloupce (zvoleno podle [22]), ve kterých je průměrné zastoupení znaků v konsenzuálním histogramu větší než uživatelem zvolená

procentuální hodnota, pro aminokyselinové sekvence je přednastavená hodnota 50 %, pro nukleotidové 75 %. Nalezení bloků je zobrazeno na Obr. 13.

Níže je uvedena ukázka výpočtu bloků. Vlivem odstranění míst, kde je velké množství mezer a neshod se zvyšují hodnoty sumy párů, sloupcového skóre i SW skóre (resp. Z-skóre).



Obr. 13: Histogram zastoupení znaku z konsenzuální sekvence v zarovnání, žlutá barva značí průměrný signál počítaný pomocí klouzavého okna délky 3 průměr je na prvním místě okna; zelená barva značí průměrný signál při prodlužování bloku tj. okno začíná na délce 3 a postupně se prodlužuje, hodnota je zobrazena na poslední pozici okna

Ukázka výpočtu blokového sloupcového skóre

Zarovnání:

```
                VT-----------------GSREIKSQQSEVTR
                --MSEDILTVAWFTAVWCGPCKTIERPMEKIAY
                EKEQKVTTIVVNIYEDGVRGCDALNSSLECLAA
                VKAAGEKLVVLDMYTQWCGPCKVIAPKYKALSE
                A-NESKTLVVVDFTASWCGPCRFIAPFFADLAK
Konsenzuální sekvence:  V-NQKETLVVVDFYAQWCGPCRTIAPQMEELAR
```

Z Obr. 13 postupně počítáme pro tři sekvence průměrnou hodnotu histogramu.

$\overline{x}_{(1-3)}$=(40+0+20)/3=20; $\overline{x}_{(2-4)}$=(0+20+20)/3=13,3; $\overline{x}_{(3-5)}$=(20+20+20)/3=20;

$\overline{x}_{(4-6)}$= (20+20+20)/3=20; $\overline{x}_{(5-7)}$=(20+20+40)/3=26,7; $\overline{x}_{(6-8)}$=(20+40+60)/3=40;

$\overline{x}_{(7-9)}$= (40+60+40)/3=46,7; $\overline{x}_{(8-10)}$=(60+40+80)/3=60 nyní můžeme blok rozšiřovat:

$\overline{x}_{(8-11)}$= (60+40+80+40)/4 =55; $\overline{x}_{(8-12)}$= (60+40+80+40+40)/5 =52;

$\overline{x}_{(8-13)}$= (60+40+80+40+40+40)/6 =50- pozice 13 již do bloku nepatří.

Blok 1 je od 8 do 12 pozice.

3.4 Výsledky testování kvality zarovnání

Pro vícenásobné zarovnání je důležité, jak již bylo dříve zmíněno, zachování důležitých struktur, funkčnosti, aj. Je to velice komplexní disciplína, proto je ruční zarovnání stále považováno za standard oproti programovému. I proto byla data pro porovnání vybírána z databází, která ruční zarovnání zahrnuje.

Pro hodnocení byly použity statistické testy Chí kvadrát test a Friedmanův test. Pomocí testu Chí kvadrát bylo zkoumáno, zda se výsledky nástrojů statisticky liší od ručních zarovnání BaliBase a RNA webu, nulová hypotéza byla stanovena následovně: Hodnoty parametrů očekávaných (ohodnocení ručních zarovnání) jsou shodné s hodnotami parametrů použitého nástroje. U žádného z nástrojů nelze vyvrátit nulovou hypotézu (na hladině 5 % statistické významnosti). Ukázka výsledků testu Chí kvadrát pro entropii datasetu RV11 je uveden v Příloze VI.

Friedmanovým testem byly hledány statisticky významné odlišnosti mezi nástroji. Nulová hypotéza (Všechny nástroje mají shodné hodnoty parametrů) opět nejde na hladině 5 % významnosti vyloučit. Ukázka výsledků Friedmanova testu pro datasetu RV30 je uvedena v Příloze VII.

V zarovnáních nebyly nalezeny statisticky významné změny, což ukazuje na to, že všechny nástroje jsou dobře sestavené a jejich vytvoření dává smysl.

Namístě je subjektivní hodnocení nástrojů použitím jednodušších statistických postupů. Toto hodnocení je provedeno s využitím normalizace (provedené pomocí vzorců (8), (9)) a průměru (funkce *mean*).

$$x_i(r) = \frac{[x_i(r) * n] + 1}{2}, \tag{8}$$

kde x_i je sloupec, r značí pořadí prvku ve sloupci, n koeficient, který je vypočítán podle vzorce (8).

$$n = \frac{1}{\sqrt{(x_i * x_i) + 1}} \tag{9}$$

Ze získaných průměrů je sestaven žebříček a následně nejlepší tři jsou uvedeny v Tab. 5. Celkově nejvhodnější nástroj je určen nejnižším součtem pořadí všech čtyř parametrů.

Tab. 5: Vyhodnocení nejlepších nástrojů

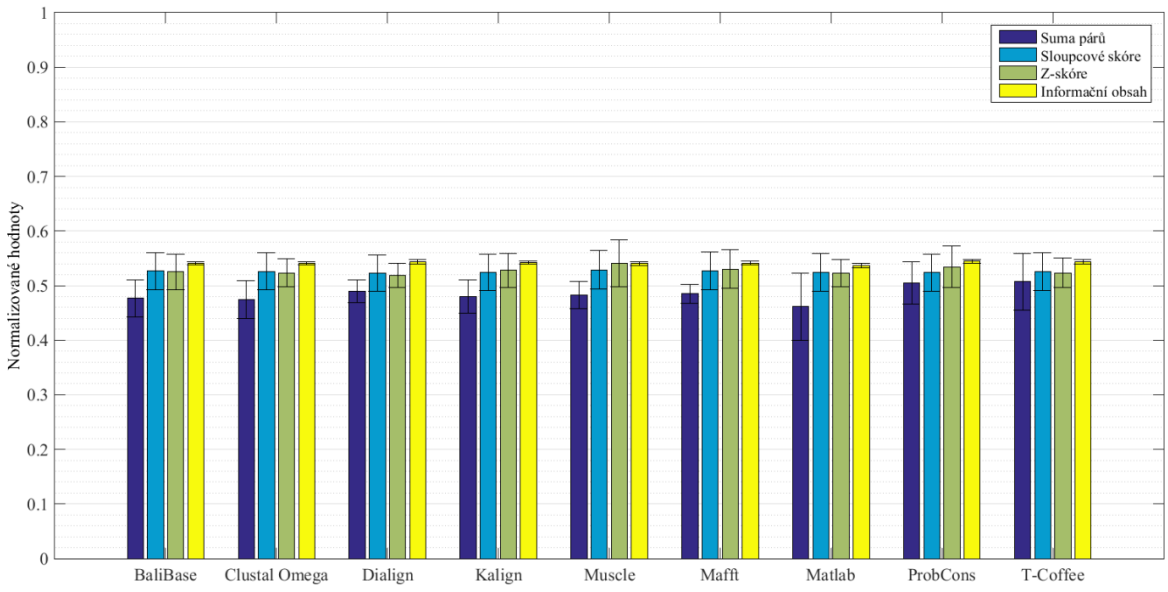
Druh zarovnání	Z hlediska sumy párů	Z hlediska sloupcového skóre	Z hlediska Z-skóre	Z hlediska entropie	Nejvhodnější celkově
RV11	Muscle Kalign Mafft	Mafft T-Coffee Muscle	Muscle Mafft Matlab	Dialign ProbCons T-Coffee	Mafft Muscle Kalign
RV12	ProbCons T-Coffee Kalign	Mafft Muscle T-Coffee	Muscle Mafft ProbCons	Dialign ProbCons Kalign	ProbCons Muscle Mafft
RV20	T-Coffee ProbCons Dialign	Mafft Muscle Matlab	Matlab Muscle Mafft	ProbCons Dialign T-Coffee	Mafft T-Coffee
RV30	ProbCons Kalign T-Coffee	Muscle Mafft Clustal	Muscle Kalign Mafft	ProbCons T-Coffee Dialign	Muscle Kalign Mafft
RV40	T-Coffee ProbCons Dialign	Muscle Mafft Clustal	Muscle Mafft Kalign	T-Coffee ProbCons Dialign	Muscle T-Coffee Mafft
RV50	T-Coffee ProbCons Muscle	Muscle Mafft Clustal	Muscle ProbCons Mafft	ProbCons T-Coffee Dialign	ProbCons T-Coffee Muscle, Mafft
rRNA	T-Coffee Clustal	Muscle Mafft	Muscle Mafft	Kalign T-Coffee	Muscle T-Coffee

3.5 Zhodnocení nástrojů

Vytvořený nástroj byl aplikován na výše popsané sady sekvencí. Pomocí získaných parametrů je možné říci, který nástroj je vhodnější.

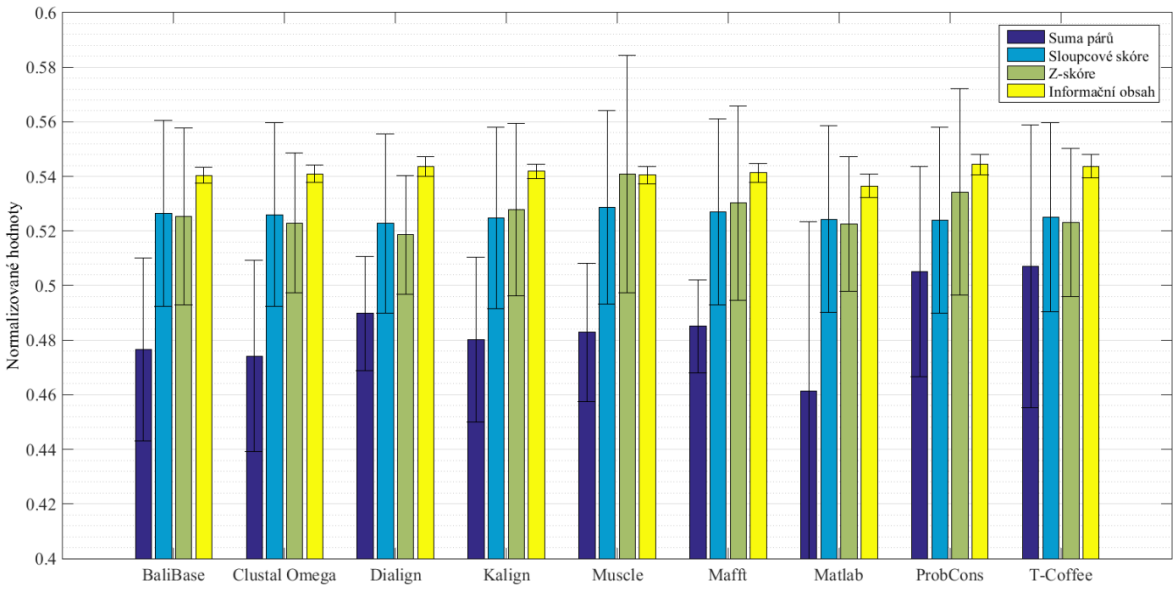
RV50

Pro sekvence s dlouhými nezarovnatelnými inzercemi jsou výsledky uvedeny v Tab. 6. a znázorněny na Obr. 14. Pro lepší přehlednost je hodnocení zobrazeno na Obr. 15, u ostatních datasetů jsou používány pouze tyto detailnější grafy. Reálné hodnoty získané programem při přednastaveném módu pro všechna zarovnání jsou vypsány v Tab. 7, Tab. 8, Tab. 9, a Tab. 10. Je patrné, že směrodatné odchylky jsou velmi podobné, což opět potvrzuje nestatistické odlišnosti nástrojů a podobnost s ručním zarovnáním.



Obr. 14: Hodnocení RV50; rozsahy znázorňují směrodatnou odchylku hodnot

V případě vícenásobného zarovnání sekvencí s dlouhými inzercemi, které nelze zarovnat, bych doporučila zvolit nástroj ProbCons, Muscle, T-Coffee nebo Mafft. V případě specifických požadavků uživatele (např. je důležité, aby byla co největší shoda ve sloupci, vyžaduje spíše lokální zarovnání,…), je doporučení uvedeno níže.



Obr. 15: Hodnocení RV50, detail

Tab. 6: Průměry normalizovaných hodnot parametrů pro dataset RV50, červené hodnoty značí nejlepší výsledek

Normalizované hodnoty	Suma párů	Sloupcové skóre	Z-skóre	Informační obsah
BaliBase	0,477	0,526	0,525	0,540
Clustal	0,474	0,526	0,523	0,541
Dialign	0,490	0,523	0,519	0,544
Kalign	0,480	0,525	0,528	0,542
Muscle	0,483	0,529	0,541	0,540
Mafft	0,485	0,527	0,530	0,541
Matlab	0,461	0,524	0,522	0,537
ProbCons	0,505	0,524	0,534	0,544
T-Coffee	0,507	0,525	0,523	0,544

Suma párů může nabývat i záporných hodnot, jak tomu je téměř ve všech případech datasetu RV50. Záporné hodnoty pro zarovnání jsou způsobeny dlouhými nezarovnatelnými inzercemi. Právě v místě, kde se nachází daná inzerce, je třeba v ostatních místech doplnit mezery. Mezery zásadně snižují skóre a dostáváme se proto do záporných hodnot.

Na Obr. 15 je vidět, že parametr sumy párů pro dataset RV50 je nejvyšší pro nástroje ProbCons a T-Coffee. Z Tab. 7 je zřejmé, že kromě T-Coffee a ProbCons je vhodný také nástroj Muscle.

Tab. 7: Hodnoty sumy párů pro zarovnání datasetu RV50, červené hodnoty značí nejlepší výsledek

Zarovnání	BaliBase	Clustal	Dialign	Kalign	Muscle	Mafft	Matlab	ProbCons	T-Coffee
50001	-1139550	-1031944	-1227319	-802000	-752367	-815366	-1045725	-916940	-1008371
50002	-261760	-336504	-385935	-240292	-193515	-217078	-267359	-296087	-268617
50003	-1245692	-1901642	-1816727	-656604	-479354	-618918	-1793308	-1026038	-1224142
50004	-936	-2008	-4049	-597	-348	-282	-3741	-410	-444
50005	24531	20912	18155	26530	28883	27402	16411	30803	27382
50006	-2878845	-3327733	-1726151	-2154006	-2101338	-1890630	-7816624	3300429	1951847
50007	-94542	-250614	-187142	-100516	-97505	-114514	-273481	-67418	-94387
50008	-157869	-205038	-281896	-158777	-148246	-168770	-204081	-145460	-168333
50009	-568790	-485933	-453070	-328431	-350624	-359331	-429861	-326799	-366927
50010	-314784	-303104	-349654	-266123	-187833	-285810	-211580	-230738	-272429
50011	-1294241	-1328701	-933766	-1266641	-1191398	-1159467	-2183484	402648	-829684
50012	-5120887	-5201090	1923139	-4795614	-3817794	-2234081	-7606770	4636005	7730875
50013	-23925	-28532	-23323	-26789	2450	-18964	-28849	-20133	-20962
50014	-34755	-179833	-213239	-51040	140803	-8245	-232628	169837	18441
50015	-1508514	-1489158	-357182	-1375820	-1329494	-1169805	-2097708	-1096078	-346453
50016	-609244	-704034	-596022	-634182	-628318	-648360	-830662	-1096078	-558369

Na Obr. 15 je patrné, že z parametrů má nejmenší rozptyl sloupcové skóre. Při prozkoumání jednotlivých řádků v Tab. 8 je zřejmé, že pro jedno konkrétní zarovnání jsou z různých nástrojů získány velmi podobné výsledky (např. zarovnání 50004 má nejnižší hodnotu nástroj Dialign 0,223 a nejvyšší společně BaliBase a T-Coffee 0,230; rozptyl je v tisícinách, respektive v desetinách procent). Pro sekvence s velkými inzercemi se nejvíce z hlediska sloupcového skóre hodí Muscle, Mafft a Clustal (jak bylo uvedeno v Tab. 5).

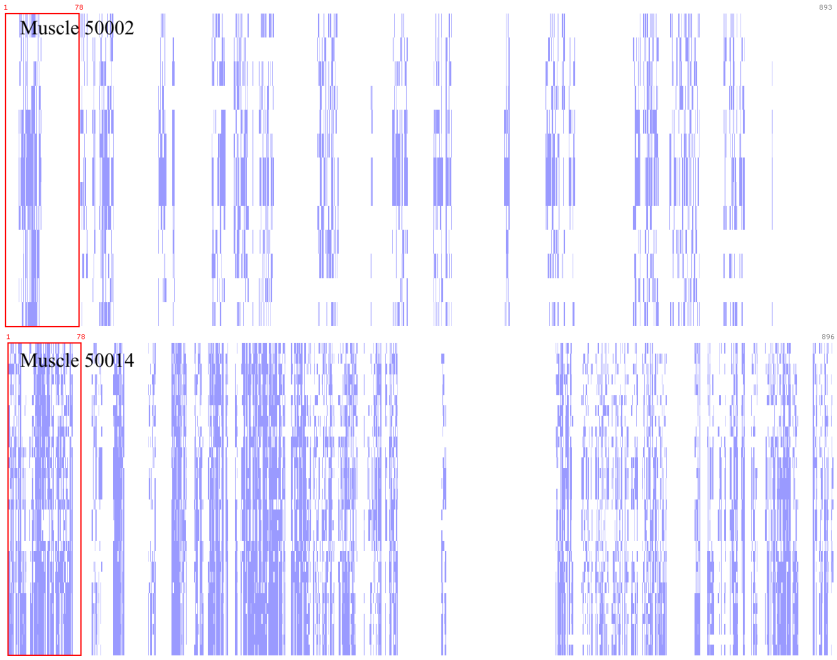
Tab. 8: Hodnoty sloupcového skóre pro zarovnání datasetu RV50, červené hodnoty značí nejlepší výsledek

Zarovnání	BaliBase	Clustal	Dialign	Kalign	Muscle	Mafft	Matlab	ProbCons	T-Coffee
50001	0,034	0,035	0,026	0,029	0,035	0,032	0,025	0,028	0,033
50002	0,008	0,007	0,003	0,005	0,008	0,009	0,000	0,005	0,006
50003	0,016	0,017	0,009	0,013	0,016	0,016	0,011	0,011	0,014
50004	0,230	0,226	0,223	0,228	0,227	0,229	0,229	0,227	0,230
50005	0,236	0,232	0,213	0,226	0,237	0,236	0,224	0,231	0,235
50006	0,012	0,012	0,008	0,011	0,013	0,012	0,008	0,007	0,008
50007	0,011	0,011	0,008	0,009	0,016	0,015	0,008	0,011	0,007
50008	0,053	0,053	0,044	0,053	0,052	0,054	0,044	0,049	0,051
50009	0,053	0,057	0,041	0,054	0,072	0,070	0,067	0,046	0,046
50010	0,022	0,023	0,015	0,021	0,025	0,025	0,012	0,018	0,021
50011	0,011	0,009	0,006	0,006	0,009	0,010	0,006	0,005	0,008
50012	0,014	0,010	0,005	0,011	0,013	0,010	0,013	0,005	0,004
50013	0,140	0,141	0,138	0,133	0,174	0,147	0,136	0,138	0,145
50014	0,084	0,081	0,058	0,072	0,107	0,080	0,082	0,066	0,073
50015	0,007	0,007	0,011	0,006	0,008	0,007	0,001	0,007	0,005
50016	0,016	0,015	0,009	0,014	0,017	0,018	0,006	0,007	0,012

SW skóre se dále přepočítává na Z-skóre, pořadí nástrojů z hlediska tohoto parametru lze určit Obr. 15 a z Tab. 9. SW skóre na rozdíl od sumy párů nepřipouští záporné hodnoty, proto můžeme v Tab. 9 vidět u setu 50002 a 50007. V těchto zarovnáních je obsaženo velké množství mezer, doplňující dlouhé nezarovnatelné inzerce. Porovnání zarovnání s nejvyšším ohodnocením a nulovým ohodnocením z datasetu RV50 je zobrazeno na Obr. 16. Nejvhodnější nástroje z hlediska Z-skóre byly určeny Muscle, ProbCons a Mafft.

Tab. 9: Hodnoty SW skóre pro zarovnání datasetu RV50, červené hodnoty značí nejlepší výsledek

Zarovnání	BaliBase	Clustal	Dialign	Kalign	Muscle	Mafft	Matlab	ProbCons	T-Coffee
50001	68801	74798	66514	75898	78485	77073	68229	75102	76075
50002	0	0	0	0	0	0	0	0	0
50003	58927	57749	59183	74711	17400	78679	60032	124126	60438
50004	19846	19812	19035	20431	20807	20849	19451	19783	19868
50005	63119	60764	57003	61428	64159	64140	59623	62755	62640
50006	2254	5138	0	3735	89955	2329	5079	94269	5
50007	0	0	0	0	0	4	0	3	0
50008	20547	14105	12243	23161	26960	22896	14502	14718	15149
50009	13266	17421	11392	38973	79132	46363	35197	41581	19078
50010	8548	12011	5660	15695	26402	13419	12122	11891	10679
50011	16487	15768	12392	12352	63255	19028	4366	13167	9114
50012	658	679	638	665	21415	655	684	5	648
50013	51050	49377	48181	52442	48423	53056	48599	49421	51132
50014	128005	80351	38580	116114	188512	139663	73528	107176	85881
50015	1481	1701	2331	1590	8750	1631	636	443	1060
50016	1124	1613	689	1517	325	1882	875	443	1039



Obr. 16: Porovnání zarovnání vytvořené pomocí nástroje Muscle s porovnáním SW skóre, nahoře set 50002 se skórem 0, dole set 50014 se skórem 188512. Zarovnání jsou zobrazena pomocí funkce multialignviewer v módu Blosum50 Score

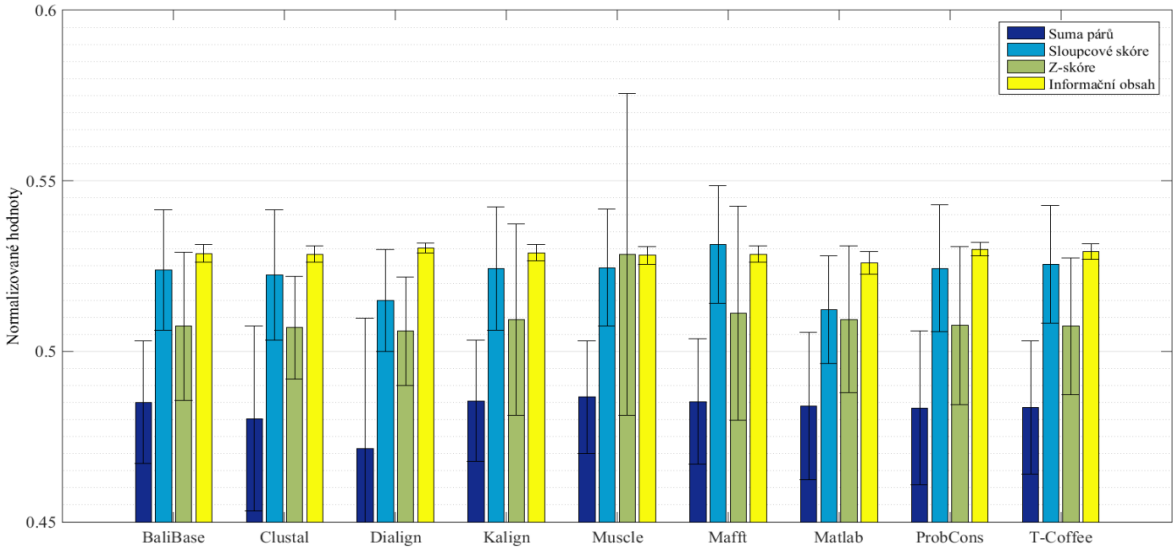
Posledním parametrem hodnocení je entropie. Aby z grafu bylo možné pro nejlepší výsledek hledat nejvyšší hodnotu je na Obr. 15 místo entropie uveden informační obsah, přepočet je možný pomocí vzorců (8) a (9). Nejnižší hodnoty entropie (viz. Tab. 10) se projevily u nástrojů ProbCons, T-Coffee a Dialign.

Tab. 10: Hodnoty entropie [bit] pro zarovnání datasetu RV50, červené hodnoty značí nejlepší výsledek

Zarovnání	BaliBase	Clustal	Dialign	Kalign	Muscle	Mafft	Matlab	ProbCons	T-Coffee
50001	1,755	1,825	1,536	1,581	1,743	1,748	2,255	1,474	1,617
50002	1,195	1,329	0,936	1,288	1,538	1,379	1,842	1,001	1,224
50003	1,200	1,179	0,940	1,005	1,118	1,064	1,477	0,884	0,957
50004	1,343	1,337	1,318	1,321	1,325	1,324	1,395	1,312	1,321
50005	1,264	1,264	1,230	1,209	1,246	1,235	1,319	1,214	1,239
50006	1,046	0,945	0,739	0,918	0,985	0,923	1,633	0,570	0,633
50007	1,631	1,652	1,212	1,330	1,475	1,550	2,228	1,194	1,345
50008	1,109	1,111	1,094	1,069	1,087	1,080	1,143	0,976	1,052
50009	1,673	1,566	1,283	1,505	1,817	1,757	2,095	1,303	1,273
50010	1,619	1,518	1,308	1,454	1,617	1,516	1,868	1,136	1,216
50011	1,210	1,117	0,829	1,085	1,133	1,092	1,361	0,575	0,829
50012	1,223	0,916	0,478	1,011	1,017	0,784	1,416	0,433	0,362
50013	1,499	1,504	1,411	1,464	1,425	1,488	1,564	1,406	1,465
50014	1,346	1,308	1,048	1,173	1,397	1,255	1,445	1,058	1,178
50015	1,138	1,053	1,248	1,030	1,155	0,962	1,375	0,933	0,619
50016	1,155	1,068	0,737	1,166	1,243	1,153	2,027	0,933	0,826

RV11

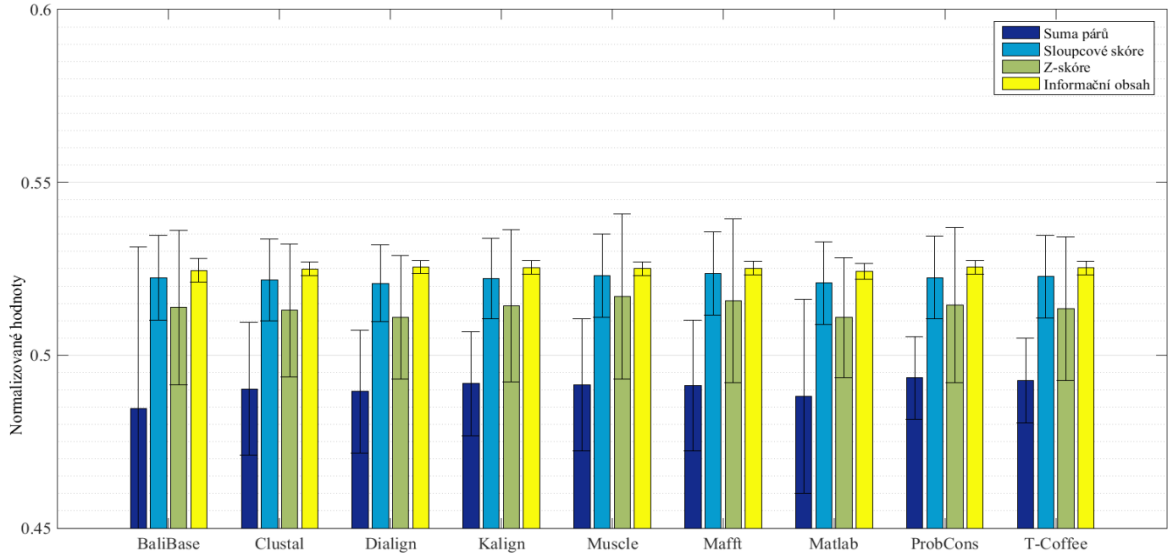
Pro sadu sekvencí, kde je shoda menší než 20 % a sekvence jsou různě dlouhé, vyšly nejlépe nástroje Mafft, Muscle a Kalign. Grafické znázornění je možné vidět na Obr. 17. Muscle byl výrazně lepší při porovnání pomocí Z-skóre, naopak Dialign byl dost horší z hlediska sumy párů.



Obr. 17: Vyhodnocení zarovnání datasetu RV11: Nahoře celý rozsah; dole detail; rozpětí udává směrodatnou odchylku dat

RV12

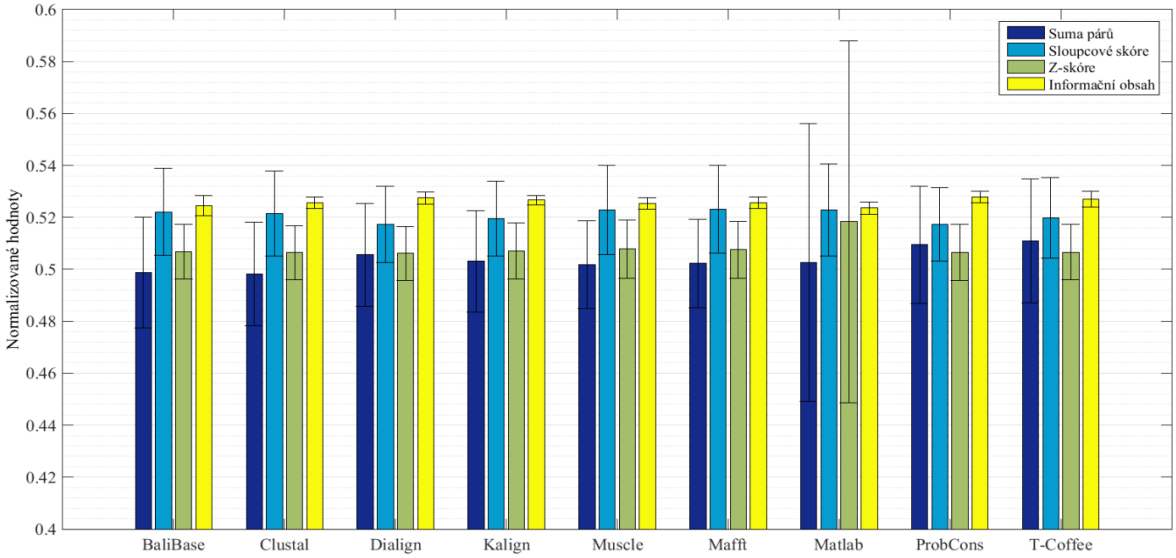
Pro sekvence z datasetu RV12 byly výsledky velice vyrovnané. Sekvence byly oproti datům ze setu RV11 více konzervované, bylo proto snazší vyhodnotit vzájemnou podobnost a zarovnat. Tento typ dat nejlépe zarovnaly nástroje ProbCons, Muscle a Mafft. Na Obr. 18 jsou výsledky hodnocení graficky znázorněny.



Obr. 18: Vyhodnocení zarovnání datasetu RV12: Vpravo celý rozsah; vlevo detail; rozpětí udává směrodatnou odchylku dat

RV20

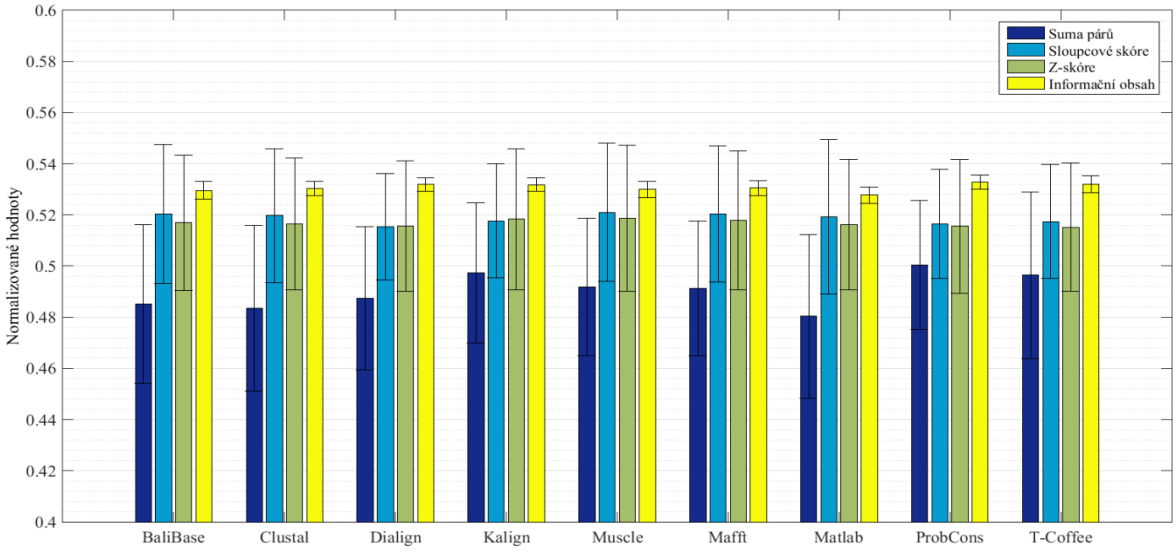
Na Obr. 19 jsou výsledky hodnocení pro sadu sekvencí RV20, velice podobných sekvencí, ke kterým je přiřazena jedna (případně více) nepodobná sekvence. Tuto úlohu nejlépe vyřešil nástroj Mafft a T-Coffee.



Obr. 19: Vyhodnocení zarovnání datasetu RV20: Vpravo celý rozsah; vlevo detail; rozpětí udává směrodatnou odchylku dat

RV30

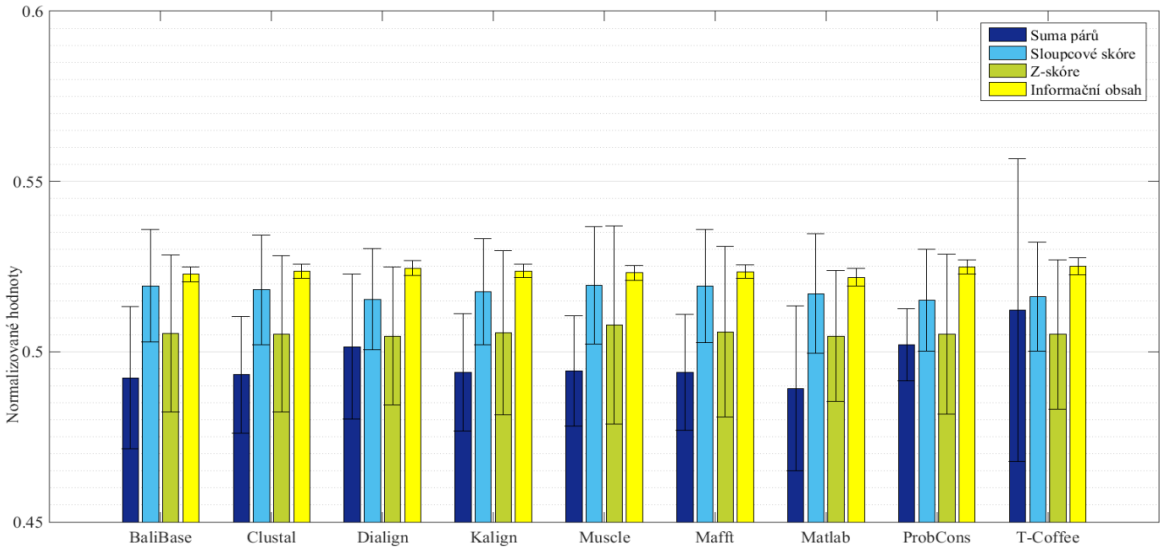
Jako nejvhodnější nástroj pro zarovnání více odlišných skupin byl určen Muscle, Kalign a Mafft (viz. Obr. 20).



Obr. 20: Vyhodnocení zarovnání datasetu RV30: Vpravo celý rozsah; vlevo detail; rozpětí udává směrodatnou odchylku dat

RV40

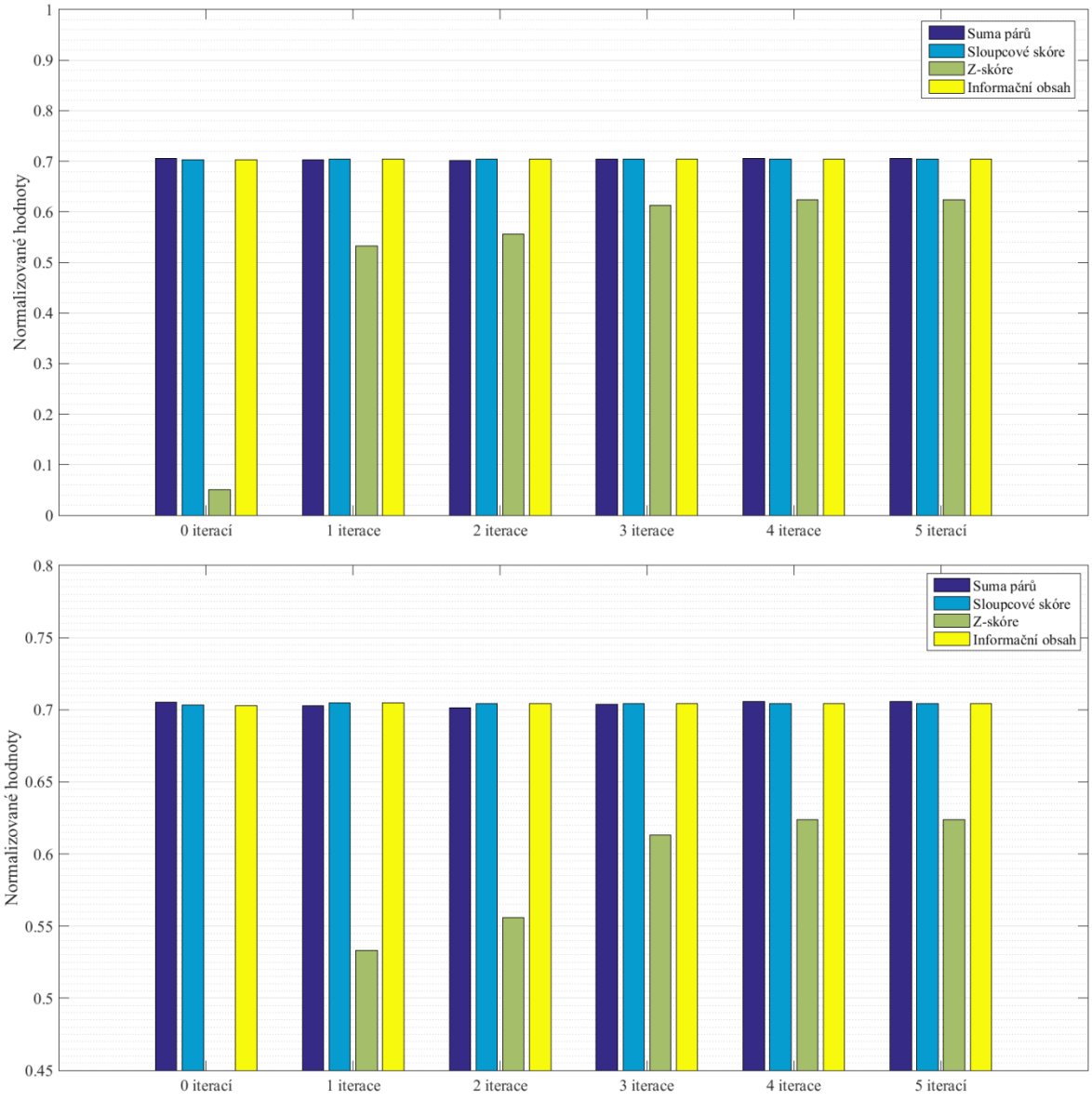
Obr. 21 ukazuje výsledky pro dataset RV40, obsahující dlouhé přesahy u některých sekvencí. Nejlépe se zarovnání těchto sekvencí podařilo nástrojům Muscle, T-Coffee a Mafft.



Obr. 21: Vyhodnocení zarovnání datasetu RV40: Vpravo celý rozsah; vlevo detail; rozpětí udává směrodatnou odchylku dat

Vliv nastavení počtu iterací nástroje Clustal na zarovnání

Clustal je jedním z nejpoužívanějších nástrojů. V jeho nastavení je možné zvolit si počet iterací pro přepočet vodícího stromu. Na Obr. 22 a v Tab. 11 je možné vidět, že nejvíce se mění Z-skóre, respektive SW skóre. Sekvence jsou v tomto případě lépe lokálně zarovnány. V případě neopakování výstavby vedoucího stromu (případ 0 iterací) je vidět v Z-skóre obrovský propad oproti iteracím. Změny ostatních parametrů jsou zanedbatelné. V tomto případě je plně dostačující opakování 4 krát. Při pěti opakováních se hodnoty parametrů nezměnily (viz. poslední dva řádky v Tab. 11).



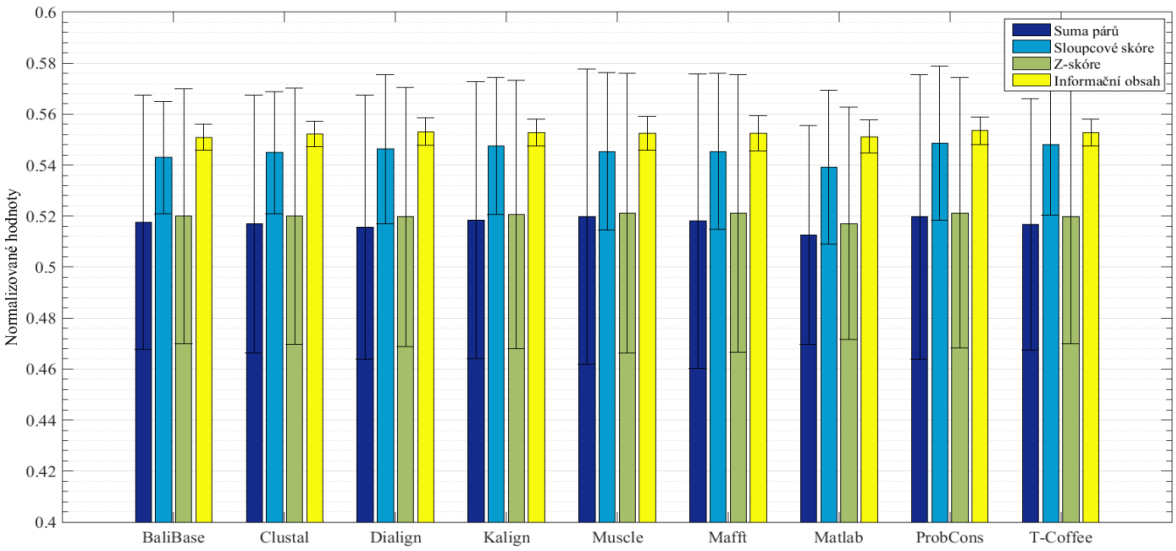
Obr. 22: Vyhodnocení zarovnání Clustal Omega při různém počtu iterací pro zarovnání BB20004

Tab. 11: Vypočítané hodnoty nástroje Clustal Omega při různém počtu iterací pro sekvenci BB20004

Počet iterací	Počet sekvencí	Délka zarovnání	Suma párů	Sloupcové skóre	SW skóre	Entropie
0	55	944	778592	0.145	1414751	1,500
1	55	932	769238	0.146	1471787	1,515
2	55	934	763735	0.146	1474469	1,513
3	55	934	772394	0.146	1481269	1,512
4	55	934	779901	0.146	1482501	1,512
5	55	934	779901	0.146	1482501	1,512

Hodnocení zarovnání pomocí sestavení bloků

V programu je možnost ohodnotit zarovnání pomocí sestavení bloku. Dataset RV40 má velké přesahy, právě pro taková zarovnání je tato funkce sestavena. Pro deset sekvencí byly vypočítány parametry z bloků, které mají průměrnou shodu minimálně 40 %. Hodnoty parametrů jsou uvedeny v Příloze V, vyobrazeny na Obr. 23.



Obr. 23: Vyhodnocení zarovnání s nastavením vytvoření bloků

Při hodnocení metodou sumy párů se u dlouhých oblastí s mezerami dostáváme k záporným hodnotám. Jak je vidět v Příloze V (Tabulka 1), při blokovém ohodnocení je suma párů častěji v kladných hodnotách.

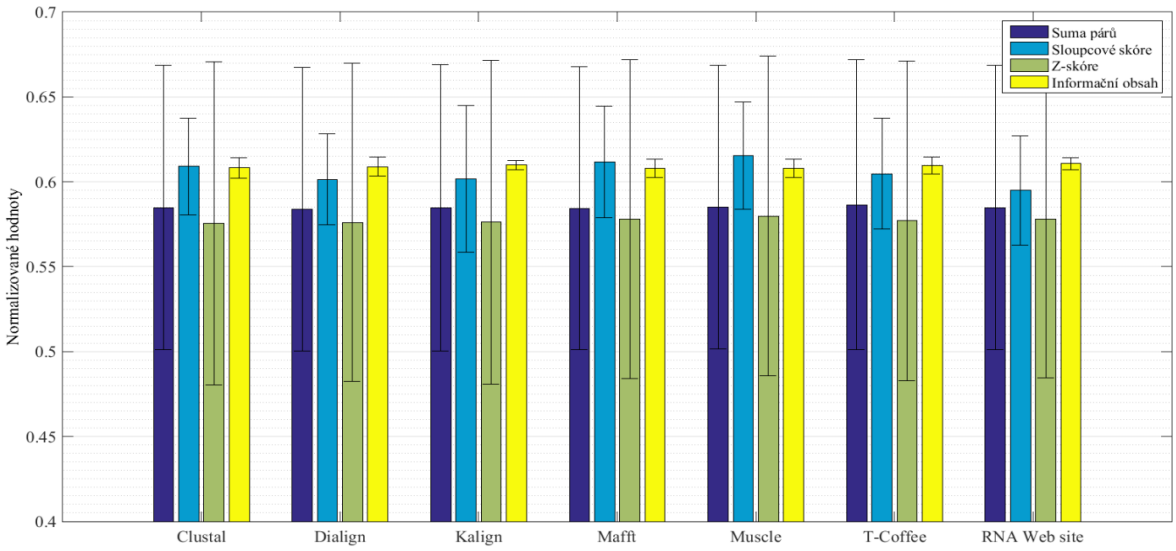
Ovlivnění sloupcového skóre při blokovém zarovnání je velice zásadní. Posun hodnot je o celý řád. Při minimální požadované shodě 40 % je ze všech zarovnání (z Přílohy V, Tabulka 2) nejlépe ohodnocen parametrem sloupcové skóre set 40010 pomocí nástroje T-Coffee hodnotou 0,388, tzn. že 38,8 % sloupců má shodu větší nebo rovnu 75 %. Hodnoty SW skóre uvedené v Příloze V (Tabulka 3) získáváme opět vyšší. Mezery opět nesnižují skóre. Celkové vyhodnocení nástrojů pomocí funkce sestavení bloků byl jako nejlepším nástrojem vyhodnocen ProbCons a Kalign.

RNA

Výsledky hodnocení zarovnání rRNA sekvencí je možné vidět na Obr. 24 a konkrétní hodoty jsou uvedeny v Tab. 12. Při zarovnávání těchto sekvencí byly nejúspěšnějšími nástroji T-Coffee a Muscle. Dále je třeba vyzdvihnout výraznější hodnoty sumy párů nástroje T-Coffee a sloupcového skóre nástroje Muscle (označeno červeně).

Tab. 12: Hodnoty parametrů nástrojů pro rRNA zarovnání, červené hodnoty značí nejlepší výsledek

	RNA Web site	Clustal	Dialign	Kalign	Muscle	Mafft	T-Coffee
Suma párů							
5S.A	2617740	2620441	2541195	2575403	2663008	2611964	2683866
5S.C	6600719	6607913	6572957	6634032	6593086	6574959	6725950
5S.M	368198	368731	367631	366402	366595	367703	371386
Sloupcové skóre							
5S.A	0,269	0,354	0,329	0,243	0,368	0,344	0,320
5S.C	0,524	0,577	0,524	0,558	0,626	0,608	0,530
5S.M	0,533	0,591	0,563	0,618	0,618	0,609	0,614
SW skóre							
5S.A	1539670	1247051	1363051	1292190	1651505	1514785	1425831
5S.C	7114429	7102853	7059398	7139518	7197577	7130370	7104418
5S.M	365215	375050	370785	376113	377291	380166	375836
Entropie							
5S.A	0,615	0,814	0,779	0,610	0,807	0,808	0,722
5S.C	0,404	0,447	0,423	0,430	0,481	0,469	0,396
5S.M	0,381	0,418	0,410	0,445	0,434	0,431	0,414



Obr. 24: Výsledky hodnocení rRNA zarovnání z RNA Web site

3.6 Celkové zhodnocení

Při porovnávání nástrojů jednotlivých sad sekvencí se v Tab. 5, shromažďující celkové výsledky, vyskytly všechny nástroje. Součty největších úspěchů a neúspěchů jsou uvedeny v Tab. 13. Při volbě nejuniverzálnějšího nástroje by byly označeny Muscle, Mafft a ProbCons. Při volbě nevhodných nástrojů by to byly Matlab (respektive implementovaná funkce *multialign*) a Dialign. Clustal, ač jeden z nejpoužívanějších nástrojů, moc dobré výsledky neměl. Tento neúspěch mohl být způsoben neoptimálním způsobem nastavení. Na Obr. 22 je např. vidět rapidní zlepšení z hlediska sumy párů při použití jedné iterace, oproti zarovnání bez iterací.

ProbCons a Dialign měly dobré výsledky z hlediska entropie. ProbCons a T-Coffee vynikaly z hlediska sumy párů. Mafft a Muscle byly nejlepší z hlediska sloupcového skóre a Muscle ještě v Z-skóre.

Tab. 13: Celkové výsledky, počty nejlepších a nejhorších umístění nástrojů

Nástroj	Suma párů		Sloupcové skóre		Z-skóre		Entropie		Celkové hodnocení	
	Nejlepší	Nejhorší	Nejlepší	Nejhorší	Nejlepší	Nejhorší	Nejlepší	Nejhorší	Nejlepší	Nejhorší
Clustal		1				1				2
Dialign		2		4		3	2			1
Kalign							1			
Muscle	1		4		6			1	3	
Mafft			3						2	
Matlab		4		1	1	2		6		4
ProbCons	2			2			3		2	
T-Coffee	4					1	1			

Při porovnání hodnot pro nukleotidové a proteinové zarovnání je zřejmé, že hodnoty nukleotidových sekvencí jsou vyšší (až na entropii, jejíž maximální hodnota pro nukleotidové sekvence je $2 \cdot \log(4)$ a pro proteinové $2 \cdot \log(24)$). Je to dáno množstvím znaků použité abecedy biologických sekvencí (4 pro nukleotidy, 24 pro proteiny). Pro sloupcové skóre je změna hodnot velice výrazná. Při ohodnocování proteinových zarovnání je hodnota skóre v setinách, kdežto u nukleotidů jsou to desetiny.

Posledním kritériem, spíše pro zajímavost, byla rychlost trvání zarovnání. Jednotlivé nástroje mají uvedené rychlostí výpočtů po splnění úkonu. U ProbCons se tento parametr bohužel nepodařil nalézt. Každým nástrojem bylo provedeno 6 zarovnání, každé zarovnání bylo zhotoveno desetkrát. Časové údaje byly dále zprůměrovány. Získané hodnoty jsou uvedené v Tab. 14. Délky zarovnání použitých k tomuto hodnocení jsou v Tab. 15. Je patrné, že nejrychlejším nástrojem byl Kalign. Naopak nejpomalejšími byly Muscle a Dialign. Tyto údaje je ale třeba brát s rezervou. Záleží na vytíženosti serveru, který zarovnání vypočítává (na denní době a počtu uživatelů, kteří mají v danou chvíli zájem o zarovnání). Zarovnání pomocí Matlabu, resp. funkce *multialign* bylo sestaveno na

počítači s procesorem Intel Core i3-3227U CPU 1,90 GHz a pamětí RAM 4,00 GB. Podle Tab. 15 můžeme označit Matlab, respektive funkci *multialign*, za nástroj, který vkládá málo mezer, nejméně prodlužuje zarovnání.

Tab. 14: Průměrná rychlost zarovnání, červené hodnoty značí nejlepší výsledek, modré nejhorší

Nástroj	Rychlost zarovnání [s]					
	11001	12001	20002	30002	40002	50001
Clustal	2,3	2,4	10,0	4,8	32,4	7,0
Dialign	1,0	2,0	12,0	22,0	89,0	27,0
Kalign	1,9	2,1	2,0	1,6	2,5	1,8
Muscle	2,0	2,2	28,9	7,1	110,4	5,6
Mafft	3,0	2,8	5,1	3,9	17,7	2,8
Matlab	1,5	1,6	6,5	8,3	28,5	9,2
ProbCons	-	-	-	-	-	-
T-Coffee	2,2	3,0	8,7	21,2	50,3	14,5

Tab. 15: Délky sekvencí použitých k hodnocení rychlosti zarovnání, červené hodnoty značí nejkratší zarovnání, modré nejdelší

Nástroj	Délka zarovnání [bp]					
	11001	12001	20002	30002	40002	50001
Clustal	96	629	815	1043	4055	878
Dialign	104	676	834	1143	8608	1095
Kalign	99	655	895	1248	3172	1010
Muscle	96	619	755	1012	1838	905
Mafft	102	626	763	1037	2308	901
Matlab	96	573	729	838	1547	727
ProbCons	96	660	1236	1219	7792	1102
T-Coffee	96	642	889	1229	12056	993

4 Závěr

V rámci této bakalářské práce byl navržen, sestaven, otestován a aplikován komplexní program testování kvality vícenásobného zarovnání. Hodnocení kvality probíhá v programu na základě 4 parametrů (sumy párů, sloupcového skóre, Z-skóre a entropie). Ohodnocení lze provézt z celých zarovnání nebo z bloků, které program vytvoří. Celkové vyhodnocení je možné sestavit individuálně či skupinově. Podobný program v české vědecké sféře není běžně dostupný. Měl by být přínosem zejména pro netechnické vědecké pracovníky (biology, genetiky, chemiky, atd.). Díky programu mohou zvolit odpovídající nástroj. Pokud uživatel například ví, že ho zajímá spíše lokální zarovnání, měl by více zohlednit Z-skóre než jiné parametry. Po volbě nástroje lze optimalizovat jeho nastavení (velikost otevření a rozšíření mezery, použití substituční matice, využití počtu iterací, aj.).

Program byl využit k hodnocení zarovnání z nástrojů Clustal Omega, Dialign, Kalign, Muscle, Mafft, ProbCons, T-Coffee a funkce multialign v Matlabu. Clustal Omega byl vybrán z důvodu vytváření zarovnání pomocí zkrácených sekvencí (tzv. emBeddingu) a z proto, že je jedním z nejpoužívanějších. Odlišnost Dialignu je vkládání mezer až na konci zarovnávání, Kalign používá progresivní algoritmus, Muscle zarovnává třístupňovým systémem. Mafft byl zvolen pro jeho možnost použití rychle Fourierovi transformace, ProbCons pro použití skrytých Markovových modelů, T-Coffee díky jeho tvorbě knihoven. Funkce multialign byla přidána z důvodu častého zpracovávání genetických dat v Matlabu. Z hlediska subjektivního pocitu při práci s nástroji nejlépe hodnotím Mafft, jeho nastavení je uživatelsky přívětivé. Uživatel seznámený s tématem se v nastavení rychle zorientuje. Parametrů k ovlivnění zarovnání má Mafft celkem osm.

Testování kvality zarovnání probíhalo pomocí databáze BaliBase a dat z porovnávacího RNA webu. Sety sekvencí z uvedených databází jsou sestaveny právě k testování vícenásobného zarovnání. Obsahují mimo jiné i ručně zarovnané standardy. BaliBase obsahuje datasety s různými sekvencemi k zarovnání. V jednom datasetu jsou obsaženy sekvence stejných úloh, který mají nástroj otestovat (např. sekvence s dlouhými inzercemi, sekvence s velkými prodlouženými, sekvence s vnitřními inzercemi aj.) Jako nejuniverzálnější nástroj byl označen Muscle a Mafft, nejhůře dopadla funkce multialign z Matlabu a paradoxně dobře na tom nebyl ani Clustal Omega, jeden z nejpoužívanějších nástrojů. Také je třeba vyzdvihnout skvělé výsledky nástroje Muscle v parametru Z-skóre, byl nejlepší šestkrát ze sedmi hodnocení. Z hlediska sumy párů získal jasné prvenství nástroj T-Coffee (čtyřikrát nejlepší ze sedmi).

Rychlost zpracování sekvencí nástrojů závisí ve většině případů na denní době, vytíženosti serveru. Při měření času dopadl nejlépe nástroj Kalign. Který ani v jednom z testovaných případů nepřesáhl jednotky sekund. Naopak nejvyšší časové nároky na zarovnání má nástroj Muscle, který v extrémním případě potřeboval více než 40 krát delší čas než nástroj Kalign.

Literatura

[1] HIGGINS, Desmond G. a Paul M. SHARP. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*. 1988, 73(1): 237-244. DOI: 10.1016/0378-1119(88)90330-7. ISSN 03781119. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/0378111988903307

[2] HIGGINS, Desmond G., Alan J. BLEASBY a Rainer FUCHS. CLUSTAL V: improved software for multiple sequence alignment. *Bioinformatics*. 1992, 8(2): 189-191. DOI: 10.1093/bioinformatics/8.2.189. ISSN 1367-4803. Dostupné také z: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/8.2.189

[3] CHENNA, R. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Researc*h. 2003, 31(13): 3497-3500. DOI: 10.1093/nar/gkg500. ISSN 1362-4962. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkg500

[4] THOMPSON, J. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research.* 1997, 25(24): 4876-4882. DOI: 10.1093/nar/25.24.4876. ISSN 13624962. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/25.24.4876

[5] SAITOU, Naruya a Masatoshi NEI. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*. 1987, 4(4). Dostupné také z: http://mbe.oxfordjournals.org/content/4/4/406.short

[6] THOMPSON, Julie D., Desmond G. HIGGINS a Toby J. GIBSON. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.*Nucleic Acids Research*. 1994, 22(22): 4673-4680. DOI: 10.1093/nar/22.22.4673. ISSN 0305-1048. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/22.22.4673

[7] SIEVERS, F., A. WILM, D. DINEEN, T. J. GIBSON, K. KARPLUS, W. LI, R. LOPEZ, H. MCWILLIAM, M. REMMERT, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology.* 2011, **7**(1): 539-539. DOI: 10.1038/msb.2011.75. ISSN 1744-4292. Dostupné také z: http://msb.embopress.org/cgi/doi/10.1038/msb.2011.75

[8] SODING, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics.* 2005, **21**(7): 951-960. DOI: 10.1093/bioinformatics/bti125. ISSN 1367-4803. Dostupné také z: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bti125

[9] NOTREDAME, Cédric, Desmond G HIGGINS a Jaap HERINGA. T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*. 2000, **302**(1): 205-217. DOI: 10.1006/jmbi.2000.4042. ISSN 00222836. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/S0022283600940427

[10] NOTREDAME, C., L. HOLM a D. G. HIGGINS. COFFEE: an objective function for multiple sequence alignments. *Bioinformatics.* 1998, 14(5): 407-422. DOI: 10.1093/bioinformatics/14.5.407. ISSN 1367-4803. Dostupné také z: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/14.5.407

[11] FENG, Da-Fei a Russell F. DOOLITTLE. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of Molecular Evolution*. 1987, 25(4): 351-360. DOI: 10.1007/BF02603120. ISSN 0022-2844. Dostupné také z: http://link.springer.com/10.1007/BF02603120

[12] KATOH, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*. 2002, **30**(14): 3059-3066. DOI: 10.1093/nar/gkf436. ISSN 13624962. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkf436

[13] KATOH, K. a D. M. STANDLEY. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*. 2013, **30**(4): 772-780. DOI: 10.1093/molbev/mst010. ISSN 0737-4038. Dostupné také z: http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/mst010

[14] NEEDLEMAN, Saul B. a Christian D. WUNSCH. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. 1970, 48(3): 443-453. DOI: 10.1016/0022-2836(70)90057-4. ISSN 00222836. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/0022283670900574

[15] EDGAR, Robert C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 2004, **5**(1). DOI: 10.1186/1471-2105-5-113. ISSN 14712105. Dostupné také z: http://www.biomedcentral.com/1471-2105/5/113

[16] EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput.*Nucleic Acids Research*. 2004, **32**(5): 1792-1797. DOI: 10.1093/nar/gkh340. ISSN 1362-4962. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkh340

[17] AL AIT, L., Z. YAMAK a B. MORGENSTERN. Dialign at GOBICS--multiple sequence alignment using various sources of external information. *Nucleic Acids Research*. 2013, 41(W1): W3-W7. DOI: 10.1093/nar/gkt283. ISSN 0305-1048. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt283

[18] MORGENSTERN, B., K. FRECH, A. DRESS a T. WERNER. Dialign: finding local similarities by multiple sequence alignment. *Bioinformatics*. 1998, 14(3): 290-294. DOI: 10.1093/bioinformatics/14.3.290. ISSN 1367-4803. Dostupné také z: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/14.3.290

[19] LASSMANN, Timo a ErikLL SONNHAMMER. Kalign – an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*. 2005, 6(1): 298-. DOI: 10.1186/1471-2105-6-298. ISSN 14712105. Dostupné také z: http://www.biomedcentral.com/1471-2105/6/298

[20] LASSMANN, T., O. FRINGS a E. L. L. SONNHAMMER. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Research*. 2009, 37(3): 858-865. DOI: 10.1093/nar/gkn1006. ISSN 0305-1048. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkn1006

[21] DO, C. B. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*. 2005, 15(2): 330-340. DOI: 10.1101/gr.2821705. ISSN 1088-9051. Dostupné také z: http://www.genome.org/cgi/doi/10.1101/gr.2821705

[22] THOMPSON, J. D., F. PLEWNIAK a O. POCH. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*. 1999, **27**(13): 2682-2690. DOI: 10.1093/nar/27.13.2682. ISSN 0305-1048. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/27.13.2682

[23] KRANE, D. E. a RAYMER M. L. Fundamental concepts of bioinformatics. San Francisco: Benjamin Cummings, 2003, xiii, 314 s. ISBN 08-053-4633-3

[24] DAYHOFF, O.M., R.M. SCHWARTZ a B.C. ORCUTT. A Model of Evolutionary: Change in Proteins. Atlas of Protein Sequence and Structure. 1987, (5): 345-352. Dostupné také z: http://people.inf.ethz.ch/zollers/compbio/Dayhoff/dayhoff1978.pdf

[25] EDDY, Sean R. Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology*. 2004, **22**(8): 1035-1036. DOI: 10.1038/nbt0804-1035. ISSN 1087-0156. Dostupné také z: http://www.nature.com/doifinder/10.1038/nbt0804-1035

[26] SCHNEIDER, Thomas D. a R.Michael STEPHENS. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*. 1990, **18**(20): 6097-6100. DOI: 10.1093/nar/18.20.6097. ISSN 0305-1048. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/18.20.6097

[27] GONNET, G., M. COHEN a S. BENNER. Exhaustive matching of the entire protein sequence database. *Science*. 1992, **256**(5062): 1443-1445. DOI: 10.1126/science.1604319. ISSN 0036-8075. Dostupné také z: http://www.sciencemag.org/cgi/doi/10.1126/science.1604319

[28] THOMPSON, Julie D., Patrice KOEHL, Raymond RIPP a Olivier POCH. BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. *Proteins: Structure, Function, and Bioinformatics*. 2005,**61**(1): 127-136. DOI: 10.1002/prot.20527. ISSN 08873585. Dostupné také z: http://doi.wiley.com/10.1002/prot.20527

[29] SCHNEIDER, Thomas D. a R.Michael STEPHENS. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*. 1990, 18(20): 6097-6100. DOI: 10.1093/nar/18.20.6097. ISSN 0305-1048. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/18.20.6097

[30] PERIS, Guillermo a Andrés MARZAL. Normalized global alignment for protein sequences. *Journal of Theoretical Biology*. 2011, **291**(22-8): 22-28. DOI: 10.1016/j.jtbi.2011.09.017. ISSN 00225193. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/S0022519311004735

[31] SMITH, T.F. a M.S. WATERMAN. Identification of common molecular subsequences. *Journal of Molecular Biology*. 1981, **147**(1): 195-197. DOI: 10.1016/0022-2836(81)90087-5. ISSN 00222836. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/0022283681900875

[32] RIZK, G., D. LAVENIER a R. CHIKHI. DSK: k-mer counting with very low memory usage.*Bioinformatics*. 2013, **29**(5): 652-653. DOI: 10.1093/bioinformatics/btt020. ISSN 1367-4803. Dostupné také z: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btt020

[33] AHOLA, Virpi, Tero AITTOKALLIO, Mauno VIHINEN a Esa UUSIPAIKKA. A statistical score for assessing the quality of multiple sequence alignments. *B*MC Bioinformatics. 2006, 7(484): -. DOI: 10.1186/147-2105-7-484. Dostupné také z: http://www.biomedcentral.com/1471-2105/7/484/

[34] CANNONE, Jamie J, Sankar SUBRAMANIAN, Murray N SCHNARE, et al. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*. 2002, 3(1), 31. DOI: 10.1186/1471-2105-3-15. ISSN 14712105. Dostupné také z: http://www.biomedcentral.com/1471-2105/3/15

[35] About Gaps In Sequence Alignments. *European Molecular Biology* Laboratory-European Bioinformatics Institute [online]. European Molecular Biology Laboratory, 2012 [cit. 2016-04-06]. Dostupné z: https://web.archive.org/web/20120620133216/http://www.ebi.ac.uk:80/help/gaps.html

[36] CALLAGHAN, Martina. Barwitherr(errors,varargin). In: *The MathWorks, Inc.* [online]. University College London: MathWorks, 1994 [cit. 2016-05-13]. Dostupné z: http://uk.mathworks.com/matlabcentral/fileexchange/30639-bar-chart-with-error-bars/content/barwitherr.m

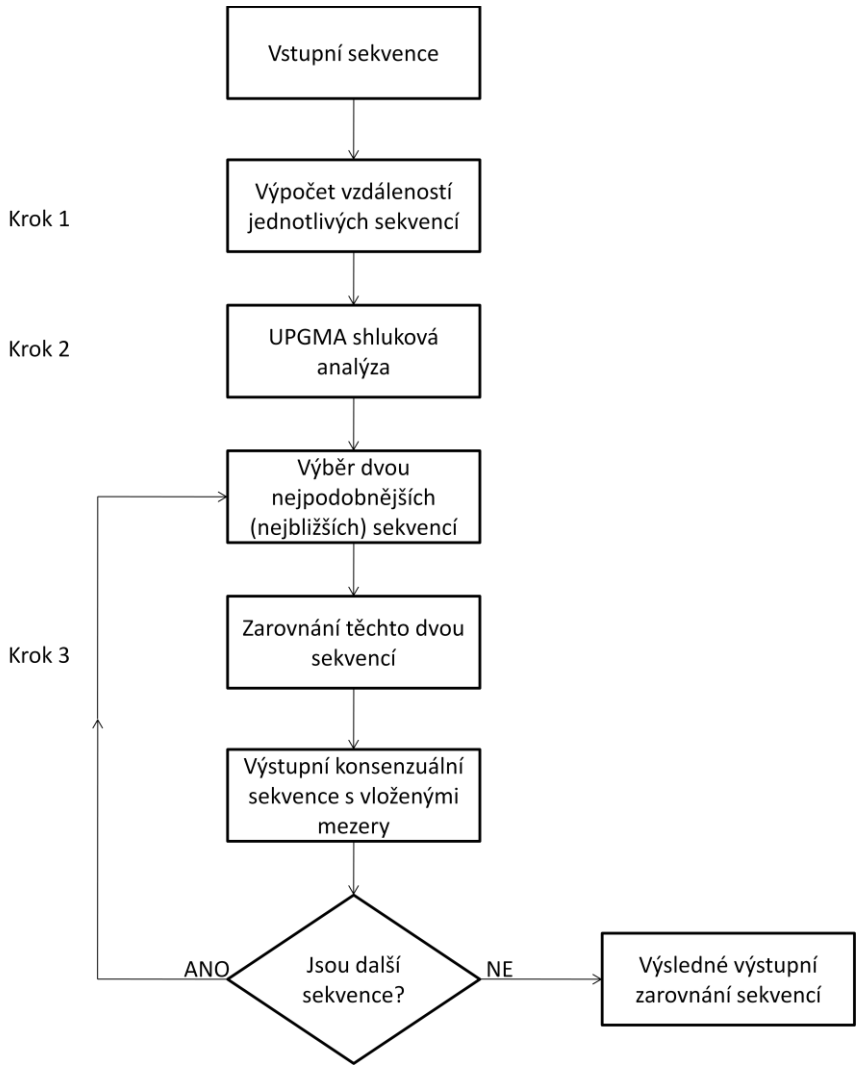
[37] Multialign: Align multiple sequences using progressive method. *The MathWorks, Inc.* [online]. The MathWorks, Inc., 1994 [cit. 2016-05-13]. Dostupné z: http://www.mathworks.com/help/bioinfo/ref/multialign.html

Seznam obrázků

Seznam tabulek

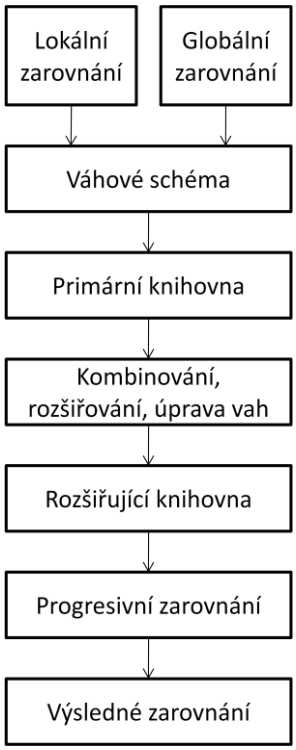
Přílohy

Příloha I



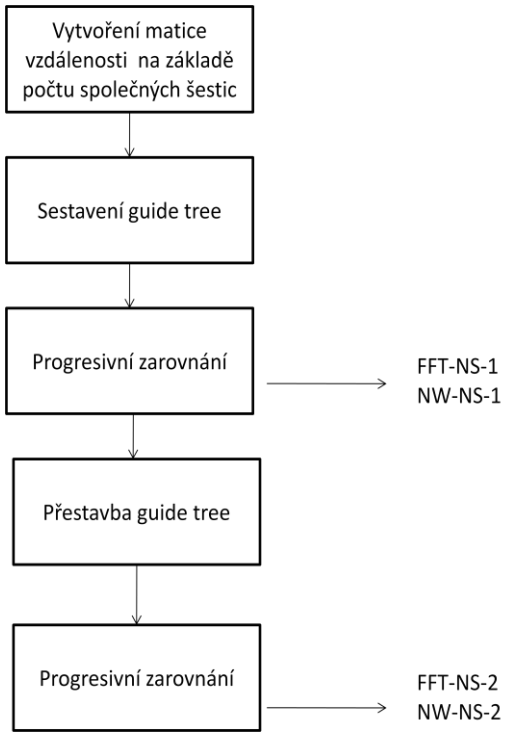
Obrázek 1: Algoritmus Clustal. Krok 1 byl původně prováděn skórováním, kladné ohodnocení shody, záporné ohodnocení neshody či mezery. Skóre bylo zaneseno do matice vzdáleností. Po sestavení této matice mohl nastat krok 2- vytvoření vodícího stromu a shlukování sekvencí pomocí metody UPGMA. Krok 3 je již to nejdůležitější, samotné zarovnávání. Vytváření konsensuálních sekvencí, vkládání mezer dokud nejsou všechny sekvence v zarovnání zahrnuty. [1]

Příloha II



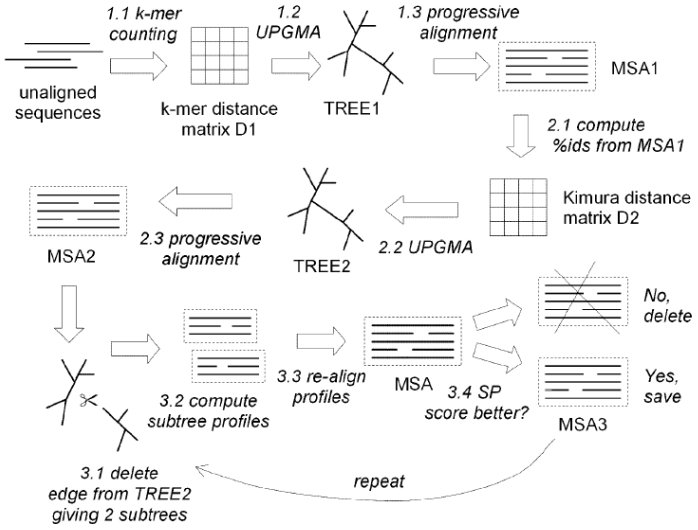
Obrázek 2: Algoritmus T-Coffee. Tento algoritmus začíná párovým zarovnáním, globálním i lokálním, každá dvojice získá váhu podle kvality zarovnání. Tyto informace jsou zahrnuté v tzv, primární knihovně, ta obsahuje již vybraná zarovnání (zarovnání se shodou větší jak 30 %). Rozšiřující knihovna obsahuje kromě zarovnání navíc informace ke vztahům k dalším sekvencím. Tyto informace jsou použity pro progresivní zarovnání - vytvoření matice distancí, sestavení vodícího stromu. Výsledné zarovnání je sestaveno. [10]

Příloha III



Obrázek 3: Mafft algoritmus. Pomocí FFT a korelace jsou sestaveny matice vzdáleností, z ní je následně metodou UPGMA sestaven vodící strom a progresivní metodou je sestaveno zarovnání. Pro lepší výsledky je možno použít varianty, které přeskládají vodící strom a zarovnání provedou ještě jednou. (Ze stránky: http://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html)

Příloha IV



Obrázek 4: Algoritmus Muscle. V průběhu jsou prováděna tři zarovnání pomocí různých metod. [16]

Příloha V

Tabulka 1: Hodnoty sumy párů pro vyhodnocení blokových oblastí

Zarovnání	BaliBase	Clustal	Dialign	Kalign	Muscle	Mafft	Matlab	ProbCons	T-Coffee
40001	-47329	4041	67582	31440	73108	59217	56939	70779	26359
40002	127450	95581	-108239	48642	41630	-68358	-77801	88362	77035
40003	33119	29901	32870	35960	32373	33971	27388	31091	32740
40004	1363154	1369940	1390716	1476929	1577620	1556996	1143273	1523877	1341284
40005	28755	26100	24159	32032	24403	24424	21567	28253	23047
40006	15160	9987	9978	14780	15542	15296	4692	15268	13494
40007	-63634	-125816	-109624	-123080	-124632	-129682	-130325	-120285	-120397
40008	18427	13994	14701	21904	21442	19943	14986	18501	14833
40009	31492	28688	15324	32048	36455	34490	17323	32152	33065
40010	4071	3525	4028	4310	4104	4434	1276	4284	3984

Tabulka 2: Hodnoty sloupcového skóre pro vyhodnocení blokových oblastí

Zarovnání	BaliBase	Clustal	Dialign	Kalign	Muscle	Mafft	Matlab	ProbCons	T-Coffee
40001	0,192	0,252	0,333	0,283	0,351	0,315	0,318	0,341	0,276
40002	0,189	0,184	0,100	0,159	0,056	0,053	0,000	0,146	0,211
40003	0,176	0,175	0,177	0,185	0,175	0,177	0,163	0,179	0,176
40004	0,172	0,181	0,189	0,191	0,198	0,186	0,172	0,198	0,188
40005	0,169	0,166	0,173	0,170	0,169	0,167	0,163	0,171	0,169
40006	0,076	0,082	0,089	0,093	0,082	0,080	0,070	0,084	0,075
40007	0,033	0,032	0,032	0,035	0,037	0,041	0,023	0,040	0,039
40008	0,119	0,123	0,111	0,135	0,119	0,123	0,108	0,121	0,115
40009	0,111	0,103	0,118	0,102	0,112	0,124	0,088	0,101	0,117
40010	0,328	0,338	0,365	0,379	0,358	0,388	0,324	0,388	0,388

Tabulka 3: Hodnoty SW skóre pro vyhodnocení blokových oblastí

Zarovnání	BaliBase	Clustal	Dialign	Kalign	Muscle	Mafft	Matlab	ProbCons	T-Coffee
40001	49091	62455	77018	68318	76154	77859	74445	78367	67456
40002	132883	104482	65713	86005	62388	49449	3310	111272	91949
40003	43651	42182	44072	46509	44347	45477	39240	44397	43914
40004	1481582	1493207	1506272	1560579	1624753	1613736	1344720	1576968	1479238
40005	41294	42622	42234	44364	44092	44270	38905	44409	42870
40006	16407	14628	15055	15980	17378	17981	9478	16900	16239
40007	6006	12618	10431	14767	11032	15160	8957	15580	15503
40008	20801	20412	20043	23145	23094	23445	18600	22345	20732
40009	34440	32924	27520	34396	39009	39983	27183	36196	35345
40010	4115	3973	4380	4560	4262	4607	3774	4684	4437

Tabulka 4: Hodnoty entropie pro vyhodnocení blokových oblastí

Entropie	BaliBase	Clustal	Dialign	Kalign	Muscle	Mafft	Matlab	ProbCons	T-Coffee
40001	2,301	1,851	1,607	1,788	1,526	1,610	1,669	1,571	1,819
40002	2,176	2,239	2,233	2,266	2,597	2,644	2,595	2,157	2,227
40003	1,668	1,658	1,640	1,622	1,651	1,624	1,708	1,653	1,658
40004	2,173	2,146	2,118	2,104	2,095	2,097	2,171	2,090	2,114
40005	1,745	1,720	1,708	1,730	1,705	1,699	1,758	1,709	1,724
40006	2,146	2,111	2,093	2,086	2,085	2,112	2,157	2,125	2,128
40007	1,973	1,770	1,733	1,776	1,793	1,764	1,881	1,729	1,732
40008	1,929	1,917	1,932	1,880	1,911	1,905	1,953	1,890	1,921
40009	2,191	2,205	2,191	2,198	2,162	2,145	2,282	2,173	2,164
40010	1,515	1,496	1,430	1,447	1,494	1,421	1,497	1,415	1,407

Příloha VI

Pozorované vs. očekávané četnosti (Entropie v RV11)
Chi-Kvadr. = 3,415632 sv = 31 p = 1,000000
POZN.: Nestejné součty pozor. a oček. četností

Případ	pozorov. Matlab	očekáv. BaliBase	P - O	(P-O)^2 /O
C: 1	1,43268	1,44830	-0,015625	0,000169
C: 2	1,61069	1,24922	0,361468	0,104592
C: 3	1,50423	1,42912	0,075112	0,003948
C: 4	1,60018	1,32395	0,276231	0,057633
C: 5	2,58591	1,98896	0,596949	0,179163
C: 6	2,23641	1,98808	0,248321	0,031017
C: 7	1,34082	1,37364	-0,032816	0,000784
C: 8	1,63878	0,93914	0,699637	0,521210
C: 9	1,57662	1,55313	0,023498	0,000356
C: 10	1,89001	1,56049	0,329525	0,069585
C: 11	1,42424	1,42521	-0,000967	0,000001
C: 12	1,73466	1,44264	0,292019	0,059110
C: 13	1,69908	1,61565	0,083435	0,004309
C: 14	1,52668	1,45357	0,073113	0,003677
C: 15	1,79797	1,32424	0,473732	0,169472
C: 16	1,54015	1,37629	0,163860	0,019509
C: 17	2,03371	1,52318	0,510528	0,171115
C: 18	1,38954	1,04169	0,347852	0,116159
C: 19	2,04484	1,81838	0,226459	0,028203
C: 20	1,57544	1,32829	0,247153	0,045987
C: 21	1,52722	1,22729	0,299930	0,073298
C: 22	0,94567	0,71767	0,228000	0,072434
C: 23	1,45183	1,23425	0,217584	0,038358
C: 24	2,18543	1,49849	0,686944	0,314912
C: 25	2,42341	1,75095	0,672452	0,258255
C: 26	2,07655	1,35417	0,722381	0,385353
C: 27	2,12224	1,42871	0,693529	0,336655
C: 28	1,93656	1,70408	0,232480	0,031716
C: 29	1,84735	1,35814	0,489205	0,176212
C: 30	2,11963	1,79228	0,327348	0,059788
C: 31	1,27610	1,05673	0,219369	0,045539
C: 32	1,58081	1,35644	0,224374	0,037114
Sčt	55,67545	45,68238	9,993078	3,415632

Pozorované vs. očekávané četnosti (Entropie v RV11)
Chi-Kvadr. = ,5400920 sv = 31 p = 1,000000
POZN.: Nestejné součty pozor. a oček. četností

Případ	pozorov. Clustal	očekáv. BaliBase	P - O	(P-O)^2 /O
C: 1	1,45154	1,44830	0,003242	0,000007
C: 2	1,09560	1,24922	-0,153623	0,018892
C: 3	1,41523	1,42912	-0,013893	0,000135
C: 4	1,33772	1,32395	0,013773	0,000143
C: 5	2,04858	1,98896	0,059620	0,001787
C: 6	1,95278	1,98808	-0,035301	0,000627
C: 7	1,22995	1,37364	-0,143691	0,015031
C: 8	1,25885	0,93914	0,319705	0,108834
C: 9	1,38646	1,55313	-0,166662	0,017884
C: 10	1,54134	1,56049	-0,019148	0,000235
C: 11	1,44861	1,42521	0,023397	0,000384
C: 12	1,49957	1,44264	0,056928	0,002246
C: 13	1,61551	1,61565	-0,000136	0,000000
C: 14	1,43245	1,45357	-0,021120	0,000307
C: 15	1,31722	1,32424	-0,007027	0,000037
C: 16	1,38015	1,37629	0,003862	0,000011
C: 17	1,57630	1,52318	0,053121	0,001853
C: 18	1,28088	1,04169	0,239194	0,054924
C: 19	1,72541	1,81838	-0,092972	0,004754
C: 20	1,30320	1,32829	-0,025091	0,000474
C: 21	1,49563	1,22729	0,268336	0,058669
C: 22	0,93249	0,71767	0,214820	0,064302
C: 23	1,40790	1,23425	0,173649	0,024431
C: 24	1,87933	1,49849	0,380846	0,096793
C: 25	1,80763	1,75095	0,056679	0,001835
C: 26	1,32677	1,35417	-0,027406	0,000555
C: 27	1,46658	1,42871	0,037869	0,001004
C: 28	1,37861	1,70408	-0,325471	0,062163
C: 29	1,37415	1,35814	0,016008	0,000189
C: 30	1,75516	1,79228	-0,037121	0,000769
C: 31	1,06951	1,05673	0,012776	0,000154
C: 32	1,32647	1,35644	-0,029971	0,000662
Sčt	46,51757	45,68238	0,835194	0,540092

Pozorované vs. očekávané četnosti (Entropie v RV11)
Chi-Kvadr. = 1,550999 sv = 31 p = 1,000000
POZN.: Nestejné součty pozor. a oček. četností

Případ	pozorov. Dialign	očekáv. BaliBase	P - O	(P-O)^2 /O
C: 1	1,33979	1,44830	-0,10851	0,008130
C: 2	0,97422	1,24922	-0,27501	0,060540
C: 3	1,29795	1,42912	-0,13117	0,012039
C: 4	1,23788	1,32395	-0,08607	0,005595
C: 5	1,41001	1,98896	-0,57895	0,168524
C: 6	1,59545	1,98808	-0,39263	0,077543
C: 7	1,17187	1,37364	-0,20177	0,029636
C: 8	1,18950	0,93914	0,25035	0,066737
C: 9	1,30027	1,55313	-0,25286	0,041166
C: 10	1,16254	1,56049	-0,39795	0,101482
C: 11	1,41925	1,42521	-0,00596	0,000025
C: 12	1,36026	1,44264	-0,08239	0,004705
C: 13	1,51709	1,61565	-0,09856	0,006013
C: 14	1,32684	1,45357	-0,12673	0,011049
C: 15	1,16698	1,32424	-0,15726	0,018675
C: 16	1,37391	1,37629	-0,00237	0,000004
C: 17	1,13278	1,52318	-0,39040	0,100063
C: 18	1,24177	1,04169	0,20009	0,038432
C: 19	1,28285	1,81838	-0,53553	0,157717
C: 20	1,29998	1,32829	-0,02831	0,000603
C: 21	1,26584	1,22729	0,03855	0,001211
C: 22	0,80276	0,71767	0,08509	0,010088
C: 23	1,26710	1,23425	0,03286	0,000875
C: 24	1,40593	1,49849	-0,09255	0,005716
C: 25	1,17319	1,75095	-0,57776	0,190645
C: 26	1,04466	1,35417	-0,30951	0,070740
C: 27	1,35751	1,42871	-0,07120	0,003549
C: 28	1,11163	1,70408	-0,59245	0,205977
C: 29	1,16589	1,35814	-0,19226	0,027216
C: 30	1,34745	1,79228	-0,44483	0,110405
C: 31	1,11437	1,05673	0,05764	0,003144
C: 32	1,22491	1,35644	-0,13153	0,012755
Sčt	40,08244	45,68238	-5,59994	1,550999

Pozorované vs. očekávané četnosti (Entropie v RV11)
Chi-Kvadr. = ,3465997 sv = 31 p = 1,000000
POZN.: Nestejné součty pozor. a oček. četností

Případ	pozorov. Kalign	očekáv. BaliBase	P - O	(P-O)^2 /O
C: 1	1,38478	1,44830	-0,063520	0,002786
C: 2	1,19981	1,24922	-0,049414	0,001955
C: 3	1,38856	1,42912	-0,040562	0,001151
C: 4	1,38817	1,32395	0,064221	0,003115
C: 5	1,95874	1,98896	-0,030225	0,000459
C: 6	1,89572	1,98808	-0,092362	0,004291
C: 7	1,20981	1,37364	-0,163824	0,019538
C: 8	0,92838	0,93914	-0,010762	0,000123
C: 9	1,39693	1,55313	-0,156198	0,015709
C: 10	1,57733	1,56049	0,016844	0,000182
C: 11	1,38268	1,42521	-0,042533	0,001269
C: 12	1,51522	1,44264	0,072577	0,003651
C: 13	1,52169	1,61565	-0,093959	0,005464
C: 14	1,40248	1,45357	-0,051090	0,001796
C: 15	1,28587	1,32424	-0,038374	0,001112
C: 16	1,41964	1,37629	0,043359	0,001366
C: 17	1,51171	1,52318	-0,011475	0,000086
C: 18	1,18180	1,04169	0,140116	0,018847
C: 19	1,60190	1,81838	-0,216484	0,025773
C: 20	1,35256	1,32829	0,024264	0,000443
C: 21	1,26352	1,22729	0,036228	0,001069
C: 22	0,89855	0,71767	0,180880	0,045589
C: 23	1,26488	1,23425	0,030639	0,000761
C: 24	1,75816	1,49849	0,259670	0,044998
C: 25	1,82446	1,75095	0,073505	0,003086
C: 26	1,38362	1,35417	0,029450	0,000640
C: 27	1,47719	1,42871	0,048482	0,001645
C: 28	1,21806	1,70408	-0,486019	0,138617
C: 29	1,33081	1,35814	-0,027337	0,000550
C: 30	1,77030	1,79228	-0,021985	0,000270
C: 31	1,05155	1,05673	-0,005177	0,000025
C: 32	1,37419	1,35644	0,017753	0,000232
Sčt	45,11907	45,68238	-0,563312	0,346600

Obrázek 5: Výsledky Chí kvadrát testu pro dataset RV11; Nahoře: vpravo Matlab, vlevo Clustal; Dole: vpravo Dialign, vlevo Kalign

Pozorované vs. očekávané četnosti (Entropie v RV11)
Chi-Kvadr. = ,4451172 sv = 31 p = 1,000000
POZN.: Nestejné součty pozor. a oček. četností

Případ	pozorov. Mafft	očekáv. BaliBase	P - O	(P-O)^2 /O
C: 1	1,31214	1,44830	-0,136160	0,012801
C: 2	1,01425	1,24922	-0,234978	0,044199
C: 3	1,40670	1,42912	-0,022423	0,000352
C: 4	1,37948	1,32395	0,055537	0,002330
C: 5	1,96296	1,98896	-0,026001	0,000340
C: 6	2,00580	1,98808	0,017713	0,000158
C: 7	1,25634	1,37364	-0,117299	0,010016
C: 8	1,14809	0,93914	0,208945	0,046487
C: 9	1,38353	1,55313	-0,169592	0,018519
C: 10	1,59276	1,56049	0,032276	0,000668
C: 11	1,37914	1,42521	-0,046066	0,001489
C: 12	1,50523	1,44264	0,062590	0,002716
C: 13	1,57238	1,61565	-0,043270	0,001159
C: 14	1,43189	1,45357	-0,021675	0,000323
C: 15	1,50021	1,32424	0,175966	0,023382
C: 16	1,34861	1,37629	-0,027677	0,000557
C: 17	1,53609	1,52318	0,012909	0,000109
C: 18	1,24709	1,04169	0,205401	0,040501
C: 19	1,68295	1,81838	-0,135427	0,010086
C: 20	1,32943	1,32829	0,001140	0,000001
C: 21	1,31857	1,22729	0,091281	0,006789
C: 22	0,87605	0,71767	0,158376	0,034950
C: 23	1,32069	1,23425	0,086440	0,006054
C: 24	1,83831	1,49849	0,339825	0,077065
C: 25	1,76449	1,75095	0,013539	0,000105
C: 26	1,64413	1,35417	0,289959	0,062087
C: 27	1,41764	1,42871	-0,011072	0,000086
C: 28	1,54223	1,70408	-0,161848	0,015372
C: 29	1,52260	1,35814	0,164459	0,019914
C: 30	1,71707	1,79228	-0,075212	0,003156
C: 31	1,07274	1,05673	0,016012	0,000243
C: 32	1,29155	1,35644	-0,064892	0,003104
Sčt	46,32115	45,68238	0,638774	0,445117

Pozorované vs. očekávané četnosti (Entropie v RV11)
Chi-Kvadr. = ,8242989 sv = 31 p = 1,000000
POZN.: Nestejné součty pozor. a oček. četností

Případ	pozorov. Muscle	očekáv. BaliBase	P - O	(P-O)^2 /O
C: 1	1,43071	1,44830	-0,017591	0,000214
C: 2	0,99483	1,24922	-0,254393	0,051805
C: 3	1,43532	1,42912	0,006201	0,000027
C: 4	1,38439	1,32395	0,060446	0,002760
C: 5	2,07772	1,98896	0,088759	0,003961
C: 6	1,91092	1,98808	-0,077163	0,002995
C: 7	1,34437	1,37364	-0,029263	0,000623
C: 8	1,31320	0,93914	0,374053	0,148982
C: 9	1,43444	1,55313	-0,118690	0,009070
C: 10	1,65804	1,56049	0,097557	0,006099
C: 11	1,38838	1,42521	-0,036831	0,000952
C: 12	1,42626	1,44264	-0,016386	0,000186
C: 13	1,59068	1,61565	-0,024962	0,000386
C: 14	1,44412	1,45357	-0,009453	0,000061
C: 15	1,59362	1,32424	0,269379	0,054797
C: 16	1,37198	1,37629	-0,004301	0,000013
C: 17	1,55507	1,52318	0,031887	0,000668
C: 18	1,21452	1,04169	0,172837	0,028677
C: 19	1,77329	1,81838	-0,045095	0,001118
C: 20	1,42040	1,32829	0,092104	0,006387
C: 21	1,38709	1,22729	0,159803	0,020808
C: 22	0,88697	0,71767	0,169305	0,039940
C: 23	1,33473	1,23425	0,100488	0,008181
C: 24	2,02633	1,49849	0,527840	0,185931
C: 25	2,00205	1,75095	0,251094	0,036008
C: 26	1,61573	1,35417	0,261558	0,050520
C: 27	1,46474	1,42871	0,036031	0,000909
C: 28	1,25812	1,70408	-0,445965	0,116711
C: 29	1,60374	1,35814	0,245598	0,044412
C: 30	1,83087	1,79228	0,038589	0,000831
C: 31	1,07280	1,05673	0,016068	0,000244
C: 32	1,36197	1,35644	0,005528	0,000023
Sčt	47,60741	45,68238	1,925030	0,824299

Pozorované vs. očekávané četnosti (Entropie v RV11)
Chi-Kvadr. = 1,112232 sv = 31 p = 1,000000
POZN.: Nestejné součty pozor. a oček. četností

Případ	pozorov. ProbCons	očekáv. BaliBase	P - O	(P-O)^2 /O
C: 1	1,43071	1,44830	-0,01759	0,000214
C: 2	0,99096	1,24922	-0,25827	0,053395
C: 3	1,39076	1,42912	-0,03836	0,001030
C: 4	1,33942	1,32395	0,01548	0,000181
C: 5	1,52616	1,98896	-0,46281	0,107690
C: 6	1,72026	1,98808	-0,26783	0,036080
C: 7	1,27550	1,37364	-0,09813	0,007011
C: 8	0,92097	0,93914	-0,01818	0,000352
C: 9	1,33033	1,55313	-0,22280	0,031960
C: 10	1,39257	1,56049	-0,16791	0,018068
C: 11	1,38779	1,42521	-0,03742	0,000982
C: 12	1,35796	1,44264	-0,08468	0,004971
C: 13	1,51707	1,61565	-0,09858	0,006014
C: 14	1,37716	1,45357	-0,07641	0,004016
C: 15	1,24148	1,32424	-0,08276	0,005172
C: 16	1,29998	1,37629	-0,07630	0,004230
C: 17	1,13376	1,52318	-0,38942	0,099562
C: 18	1,26173	1,04169	0,22004	0,046479
C: 19	1,64374	1,81838	-0,17464	0,016772
C: 20	1,26664	1,32829	-0,06165	0,002862
C: 21	1,50135	1,22729	0,27406	0,061200
C: 22	0,72085	0,71767	0,00318	0,000014
C: 23	1,25664	1,23425	0,02240	0,000406
C: 24	1,27662	1,49849	-0,22187	0,032849
C: 25	1,27082	1,75095	-0,48014	0,131660
C: 26	1,14108	1,35417	-0,21309	0,033532
C: 27	1,30972	1,42871	-0,11899	0,009910
C: 28	0,99433	1,70408	-0,70975	0,295609
C: 29	1,30036	1,35814	-0,05779	0,002459
C: 30	1,40875	1,79228	-0,38353	0,082072
C: 31	1,06023	1,05673	0,00350	0,000012
C: 32	1,21160	1,35644	-0,14484	0,015467
Sčt	41,25730	45,68238	-4,42508	1,112232

Pozorované vs. očekávané četnosti (Entropie v RV11)
Chi-Kvadr. = ,5907159 sv = 31 p = 1,000000
POZN.: Nestejné součty pozor. a oček. četností

Případ	pozorov. T-Coffee	očekáv. BaliBase	P - O	(P-O)^2 /O
C: 1	1,43592	1,44830	-0,01238	0,000106
C: 2	0,99222	1,24922	-0,25700	0,052872
C: 3	1,39815	1,42912	-0,03097	0,000671
C: 4	1,31863	1,32395	-0,00532	0,000021
C: 5	1,85028	1,98896	-0,13868	0,009670
C: 6	1,92128	1,98808	-0,06680	0,002245
C: 7	1,28885	1,37364	-0,08479	0,005233
C: 8	0,92683	0,93914	-0,01231	0,000161
C: 9	1,36412	1,55313	-0,18901	0,023001
C: 10	1,54244	1,56049	-0,01805	0,000209
C: 11	1,40431	1,42521	-0,02090	0,000306
C: 12	1,54093	1,44264	0,09829	0,006696
C: 13	1,56642	1,61565	-0,04922	0,001500
C: 14	1,40894	1,45357	-0,04463	0,001370
C: 15	1,28227	1,32424	-0,04197	0,001330
C: 16	1,38822	1,37629	0,01193	0,000103
C: 17	1,32525	1,52318	-0,19793	0,025720
C: 18	1,24930	1,04169	0,20761	0,041378
C: 19	1,64453	1,81838	-0,17385	0,016621
C: 20	1,27276	1,32829	-0,05553	0,002322
C: 21	1,43411	1,22729	0,20682	0,034853
C: 22	0,81242	0,71767	0,09475	0,012509
C: 23	1,24570	1,23425	0,01146	0,000106
C: 24	1,84377	1,49849	0,34528	0,079561
C: 25	1,49625	1,75095	-0,25471	0,037052
C: 26	1,26274	1,35417	-0,09143	0,006173
C: 27	1,35178	1,42871	-0,07693	0,004142
C: 28	1,14072	1,70408	-0,56337	0,186248
C: 29	1,29500	1,35814	-0,06314	0,002936
C: 30	1,58063	1,79228	-0,21166	0,024996
C: 31	1,06400	1,05673	0,00727	0,000050
C: 32	1,23679	1,35644	-0,11965	0,010554
Sčt	43,88556	45,68238	-1,79682	0,590716

Obrázek 6: Výsledky Chí kvadrát testu pro dataset RV11; Nahoře: vpravo Mafft, vlevo Muscle; Dole: vpravo ProbCons, vlevo T-Coffee

Příloha VII

Friedmanova ANOVA a Kendallův koeficient shody (Entropie v RV30)
ANOVA chí-kv. (N = 30, sv = 7) = 148,2444 p =0,00000
Koeficient shody = ,70593 Prům.hods. r = ,69579

Proměnná	Průměrné pořadí	Součet pořadí	Průměr	**Sm.Odch.**
Clustal	3,500000	105,0000	0,719376	0,066397
T-coffee	6,000000	180,0000	0,757502	0,077138
Dialign	5,733333	172,0000	0,753778	0,061902
Kalign	5,833333	175,0000	0,750574	0,060729
ProbCons	7,400000	222,0000	0,777419	0,064176
Mafft	3,633333	109,0000	0,722358	0,066049
Muscle	2,633333	79,0000	0,710571	0,074938
Matlab	1,266667	38,0000	0,668928	0,097003

Friedmanova ANOVA a Kendallův koeficient shody (Sloupcové sk. v RV30)
ANOVA chí-kv. (N = 29, sv = 7) = 81,13793 p = ,00000
Koeficient shody = ,39969 Prům.hods. r = ,37825

Proměnná	Průměrné pořadí	Součet pořadí	Průměr	Sm.Odch.
Clustal	5,482759	159,0000	0,038956	0,051957
T-coffee	4,172414	121,0000	0,034521	0,044309
Dialign	2,206897	64,0000	0,030625	0,041197
Kalign	3,482759	101,0000	0,034982	0,044138
ProbCons	3,482759	101,0000	0,032600	0,042079
Mafft	6,275862	182,0000	0,040545	0,052678
Muscle	6,758621	196,0000	0,041685	0,053742
Matlab	4,137931	120,0000	0,037889	0,059830

Friedmanova ANOVA a Kendallův koeficient shody (Suma párů v RV30)
ANOVA chí-kv. (N = 29, sv = 7) = 101,1034 p = ,00000
Koeficient shody = ,49805 Prům.hods. r = ,48012

Proměnná	Průměrné pořadí	Součet pořadí	Průměr	Sm.Odch.
Clustal	2,241379	65,0000	-1540609	3017346
T-coffee	5,896552	171,0000	-334510	3042926
Dialign	3,275862	95,0000	-1171304	2611100
Kalign	5,896552	171,0000	-246612	2560867
ProbCons	6,620690	192,0000	45562	2354022
Mafft	4,724138	137,0000	-813804	2445997
Muscle	5,172414	150,0000	-768296	2512049
Matlab	2,172414	63,0000	-1826827	2985322

Friedmanova ANOVA a Kendallův koeficient shody (Z-skóre v RV30)
ANOVA chí-kv. (N = 29, sv = 7) = 38,88506 p = ,00000
Koeficient shody = ,19155 Prům.hods. r = ,16268

Proměnná	**Průměrné pořadí**	Součet pořadí	Průměr	Sm.Odch.
Clustal	4,517241	131,0000	0,000000	1,01770
T-coffee	5,931034	172,0000	0,000000	0,10126
Dialign	5,241379	152,0000	-0,000000	1,01770
Kalign	3,206897	93,0000	0,000000	1,01770
ProbCons	5,586207	162,0000	-0,000000	1,01770
Mafft	2,931034	85,0000	0,000000	10,20757
Muscle	4,137931	120,0000	0,000000	1,01770
Matlab	4,448276	129,0000	0,000000	1,01770

Obrázek 7: Výsledky Friedmanova testu datasetu RV30 z programu Statistika

Příloha VIII

Uživatelská příručka programu
Nástroj testování kvality vícenásobného zarovnání

Úvod

Příručka, k nástroji pro hodnocení vícenásobného zarovnání má ukázat uživateli, jak s programem pracovat, k čemu se dá využít.

Nástroj pro testování kvality je sestaven v uživatelském rozhraní GUI Matlab. Jeho hlavním cílem je porovnat zarovnání z více hledisek.

Program umožňuje dvě formy hodnocení zarovnání.
1. Uživatel potřebuje zvolit jedno zarovnání z několika. Program zvolí , které zarovnání je podle hledisek (viz. níže) nejlepší.
2. Uživatel potřebuje zjistit, jaký nástroj/nastavení/metodu použít k zarovnávání podobných typů dat. Toto hodnocení se vyhodnocuje z více zarovnání z různých nástrojů.

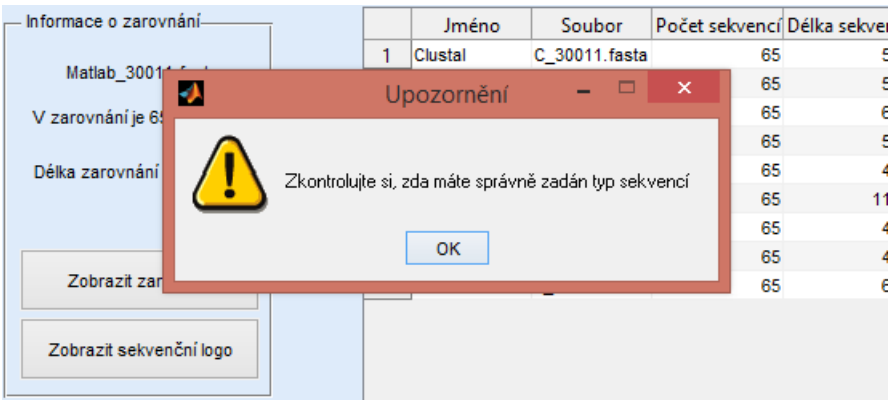
Vícenásobné zarovnání je velice obtížná disciplína, uživatel musí rozhodnou, který z parametrů je pro něj prioritou. Po výpočtu nástroj zhodnotí a řekne které zarovnání/skupina zarovnání je z daného hlediska nejvhodnější. V programu jsou uvedeny 4 parametry.
1. Suma párů
2. Sloupcové skóre
3. Z-skóre (respektive Smith-Watermanovo skóre, dále SW skóre)
4. Entropie

Popis programu

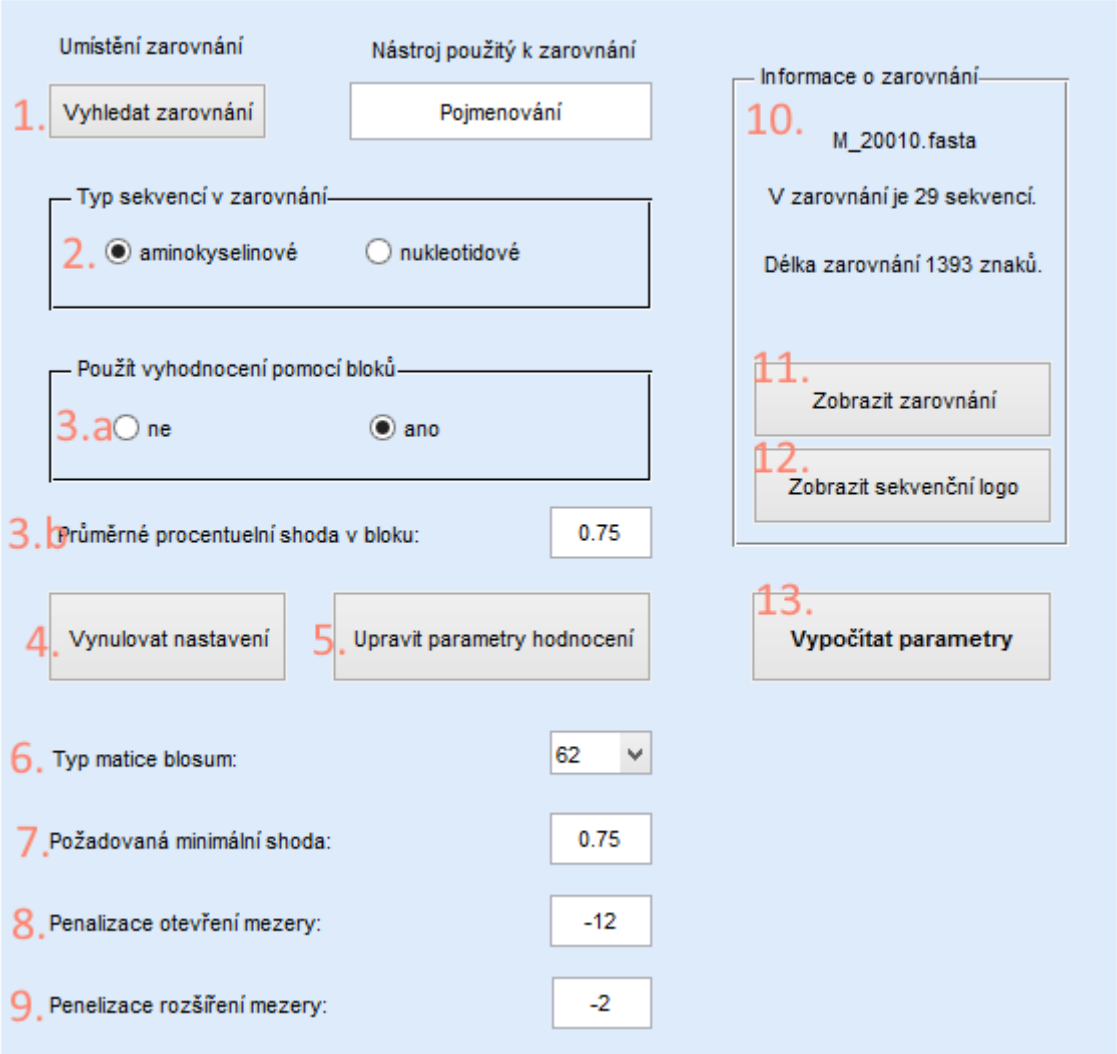
Program lze otevřít spuštěním skriptu Nastroj_testovani_kvality.m pomocí Matlabu12b a vyšší verze. Po spuštění se zobrazí hlavní okno programu. Část tohoto okna je na Obr. 2. Čísla uvedená dále se vztahují k číslům právě na tomto obrázku.

1. Tlačítko pro nahání zarovnání. Vstupy jsou požadovány ve formátu fasta a multi sequence formát(.fasta, .fa, .msf). Je možné vybrat více souborů najednou.

2. Program automaticky hodnotí, jaký typ sekvence byl vložen, v případě nesouladu s konkrétním výběrem upozorní uživatele, aby si zkontroloval správnost volby, viz. Obr. 1. Program respektuje to, co uživatel ponechá nastavené.



Obr. 1: Upozornění správnosti nastavení



Obr. 2: Nastavení vyhodnocování

3.a) Výše zmíněné parametry lze hodnotit pouze pro bloky zarovnání. Blok je taková část zarovnání, kde je shoda ve všech řádcích vyšší než X.
b) X si uživatel může zvolit. Pro nukleotidové sekvence je doporučeno vyšších hodnot (0.75, tj. 75 %) a pro aminokyselinové nižší (0.50, tj. 50 %).

4. Po stisknutí tlačítka jsou veškeré parametry změněny do přednastavených hodnot.

5. Po zapnutí programu jsou parametry na vyhodnocení skryty. Po stisku jsou pole pro úpravu parametrů zobrazeny a je možné je upravit

6. Možnost výběru typu matice Blosum pro proteinové zarovnání, přednastavená matice je Blosum62. V případě, že uživatel zarovnává velmi blízké sekvence, je doporučeno volit vyšší matice (např. blosum 90), naopak pro sekvence vzdálenější matice s nižší číselnou hodnotou (např. blosum 45). Pro nukleotidové je přednastavena substituční matice NUC44.

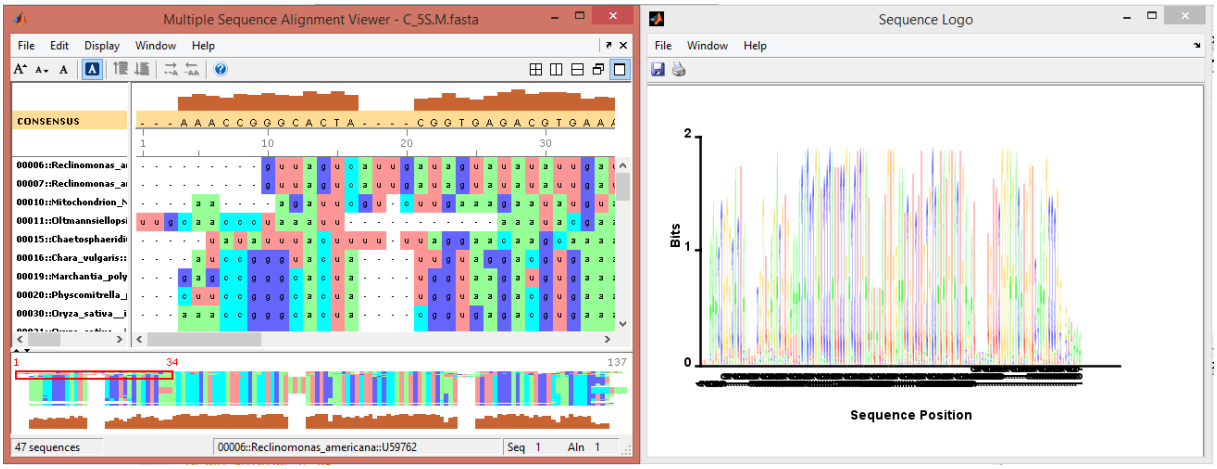
7. Požadovaná minimální shoda se týká výpočtu sloupcového skóre. Úpravou procentuelního požadavku je sloupcové skóre počítáno. Nastavuje se hodnota od 0 do 1.

8. Možnost změny hodnocení nastavení hodnoty pro otevření mezery při výpočtu Z-skóre a sumy párů (pro sumu párů je třeba hodnota penalizace mezery, tj. penalizace mezery= otevření+zavření mezery).
9. Možnost změny hodnocení nastavení hodnoty pro prodloužení mezery při výpočtu Z-skóre a sumy párů (pro sumu párů je třeba hodnota penalizace mezery, tj. penalizace mezery= otevření+zavření mezery).

10. Po nahrání sekvencí, jsou v panelu zobrazeny informace o zarovnání: název nahraného souboru, počet sekvencí v zarovnání a počet délky zarovnání.

11. Nahrané zarovnání je možné zobrazit pomocí multialignviewer, viz. Obr. 3. Při nahrání více zarovnání je třeba ze seznamu zvolit jedno konkrétní, které chce uživatel prohlížet.

12. Nahrané zarovnání je možné zobrazit pomocí sekvenčního loga, viz. Obr. 3. Je třeba zvolit úsek zájmu, ten bude zobrazen pomocí sekvenčního loga. Při nahrání více zarovnání je třeba zvolit ze seznamu jedno konkrétní, které chce uživatel zobrazit.



Obr. 3: Ukázka zobrazení zarovnání: vlevo prohlížení zarovnání, vpravo sekvenční logo

13. Po stisknutí tlačítka jsou vypočteny hodnoty sumy párů, sloupcového skóre, SW skóre a entropie. Informace o zarovnání a hodnoty parametrů jsou uloženy do tabulky.

Nahrané zarovnání a jeho parametry jsou uloženy do tabulky. Je možné s nimi dále pracovat. Na Obr. 4 je zobrazená druhá část hlavního okna. Další očíslované položky se vztahují k tomuto obrázku.

14. Místo, kde jsou zobrazovány uložené hodnoty zarovnání.
    Jméno: Jediné pole, které lze dodatečně změnit. Pro vyhodnocení výše zmíněnou první formou porovnání je jméno zásadní označení zarovnání. Pro druhou formu porovnání je nedůležité.
    Název: Uložení názvu nahraného souboru.
    Počet sekvencí: Při první formě porovnání může sloužit jako kontrola, jestli zarovnání mají stejný počet sekvencí.
    Délka sekvencí: Méně podstatná kolonka. Můžeme odhadnout, který nástroj vkládá hodně mezer, prodlužuje zarovnání.
    Suma párů: Ohodnocení zarovnání nastavenými parametry. Může nabývat libovolných hodnot (záporných, nuly i kladných). Hodnotu sumy párů vyžadujeme co nejvyšší, co nejvíce shod. Penalizační aparát má uživatel možnost volit (viz. Obr. 2, číslo 6, 8 a 9). Může zvolit substituční matici i penalizaci mezer.

Sloupcové skóre: Požadujeme co nejvyšší hodnotu. Udává kolik procent sloupců má zastoupení znaků vyšší než námi požadovaná minimální shoda (viz. Obr. 2, číslo 7). Může nabývat hodnot od 0 do 1.

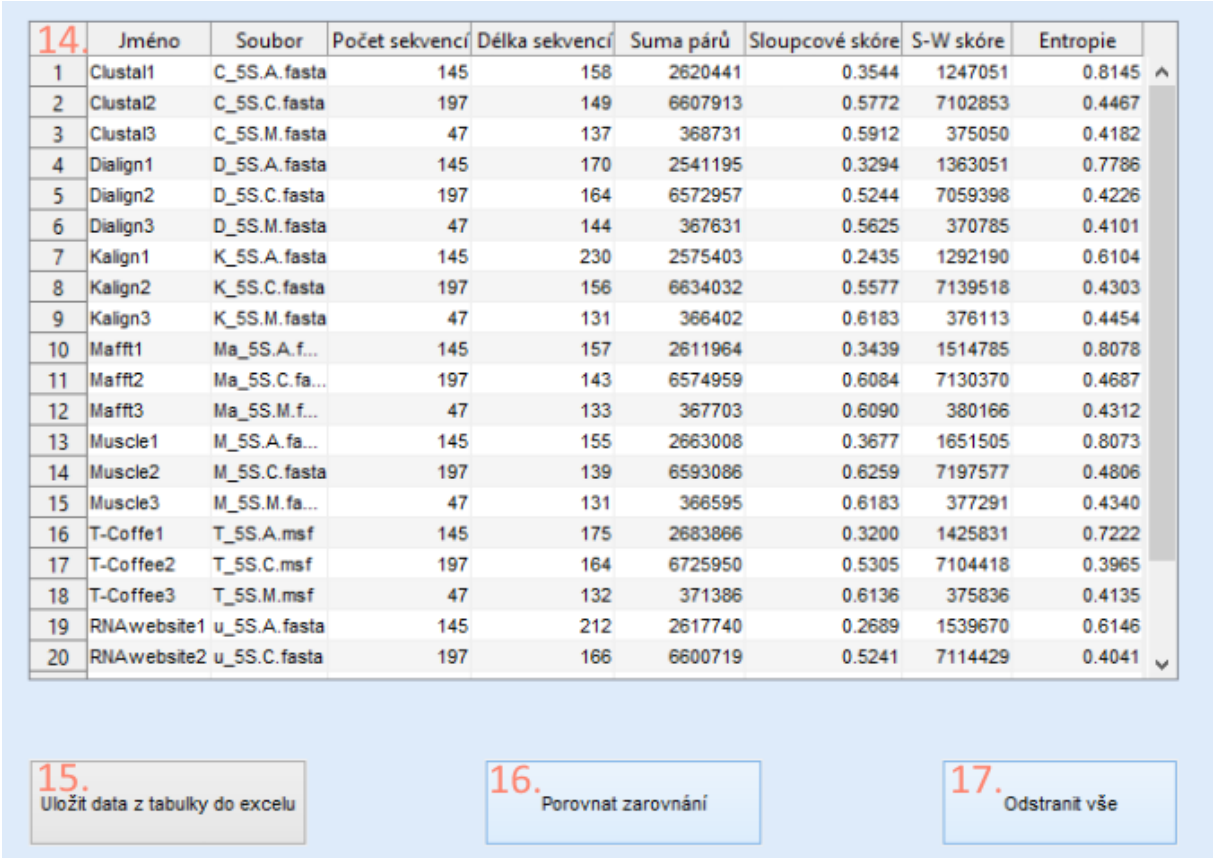
SW skóre: Požadujeme co nejvyšší hodnotu. Může nabývat libovolných hodnot od 0 výše. Je ovlivněná penalizačním aparátem jako suma párů. Je ovlivněno penalizačním aparátem (zvolenou substituční maticí a penalizací mezer).

Entropie: Jediný parametr, který chceme co nejnižší. Míra neuspořádanosti systému. Pro nukleotidy může nabývat hodnot od 0 do 2·log(4) a pro proteiny od 0 do 2·log(24).

15. Po zmáčknutí je obsah tabulky uložen ve formátu .xls na uživatelem zvolené místo.

16. Jednoduché vyhodnocení vložených zarovnání. Uživatel je informován o tom, které zarovnání je z hlediska jakého parametru nejlepší. Dále je zobrazeno grafické znázornění získaných parametrů hodnocení. Viz. dále.

17. Po zmáčknutí jsou odstraněna všechna data z tabulky. Před úkonem je uživatel ještě tázán, zda si opravdu přeje data odstranit.



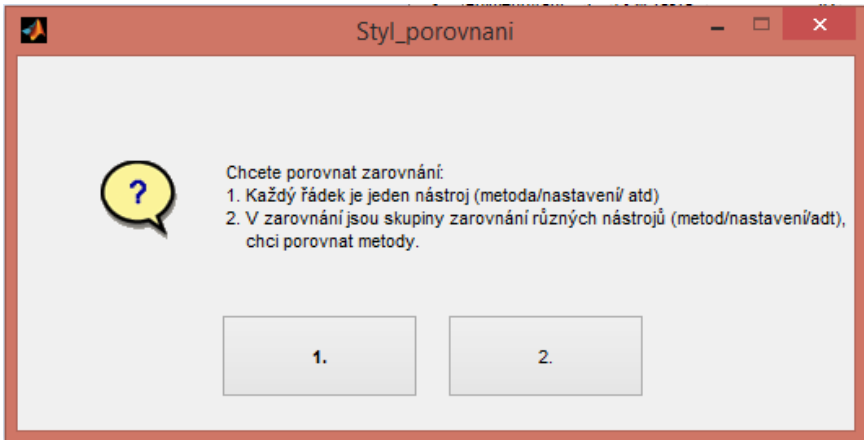
	Jméno	Soubor	Počet sekvencí	Délka sekvencí	Suma párů	Sloupcové skóre	S-W skóre	Entropie
1	Clustal1	C_5S.A.fasta	145	158	2620441	0.3544	1247051	0.8145
2	Clustal2	C_5S.C.fasta	197	149	6607913	0.5772	7102853	0.4467
3	Clustal3	C_5S.M.fasta	47	137	368731	0.5912	375050	0.4182
4	Dialign1	D_5S.A.fasta	145	170	2541195	0.3294	1363051	0.7786
5	Dialign2	D_5S.C.fasta	197	164	6572957	0.5244	7059398	0.4226
6	Dialign3	D_5S.M.fasta	47	144	367631	0.5625	370785	0.4101
7	Kalign1	K_5S.A.fasta	145	230	2575403	0.2435	1292190	0.6104
8	Kalign2	K_5S.C.fasta	197	156	6634032	0.5577	7139518	0.4303
9	Kalign3	K_5S.M.fasta	47	131	366402	0.6183	376113	0.4454
10	Mafft1	Ma_5S.A.f...	145	157	2611964	0.3439	1514785	0.8078
11	Mafft2	Ma_5S.C.fa...	197	143	6574959	0.6084	7130370	0.4687
12	Mafft3	Ma_5S.M.f...	47	133	367703	0.6090	380166	0.4312
13	Muscle1	M_5S.A.fa...	145	155	2663008	0.3677	1651505	0.8073
14	Muscle2	M_5S.C.fasta	197	139	6593086	0.6259	7197577	0.4806
15	Muscle3	M_5S.M.fa...	47	131	366595	0.6183	377291	0.4340
16	T-Coffe1	T_5S.A.msf	145	175	2683866	0.3200	1425831	0.7222
17	T-Coffee2	T_5S.C.msf	197	164	6725950	0.5305	7104418	0.3965
18	T-Coffee3	T_5S.M.msf	47	132	371386	0.6136	375836	0.4135
19	RNAwebsite1	u_5S.A.fasta	145	212	2617740	0.2689	1539670	0.6146
20	RNAwebsite2	u_5S.C.fasta	197	166	6600719	0.5241	7114429	0.4041

Obr. 4: Druhá část hlavního okna programu

Porovnání zarovnání

Podle potřeby může uživatel vyhodnoti zarovnání :

1. Jaké zarovnání je pro sekvence nejlepší. Jeden řádek je jedna položka porovnávání.
2. Porovnávání nastavení/nástroje na skupině sekvencí.

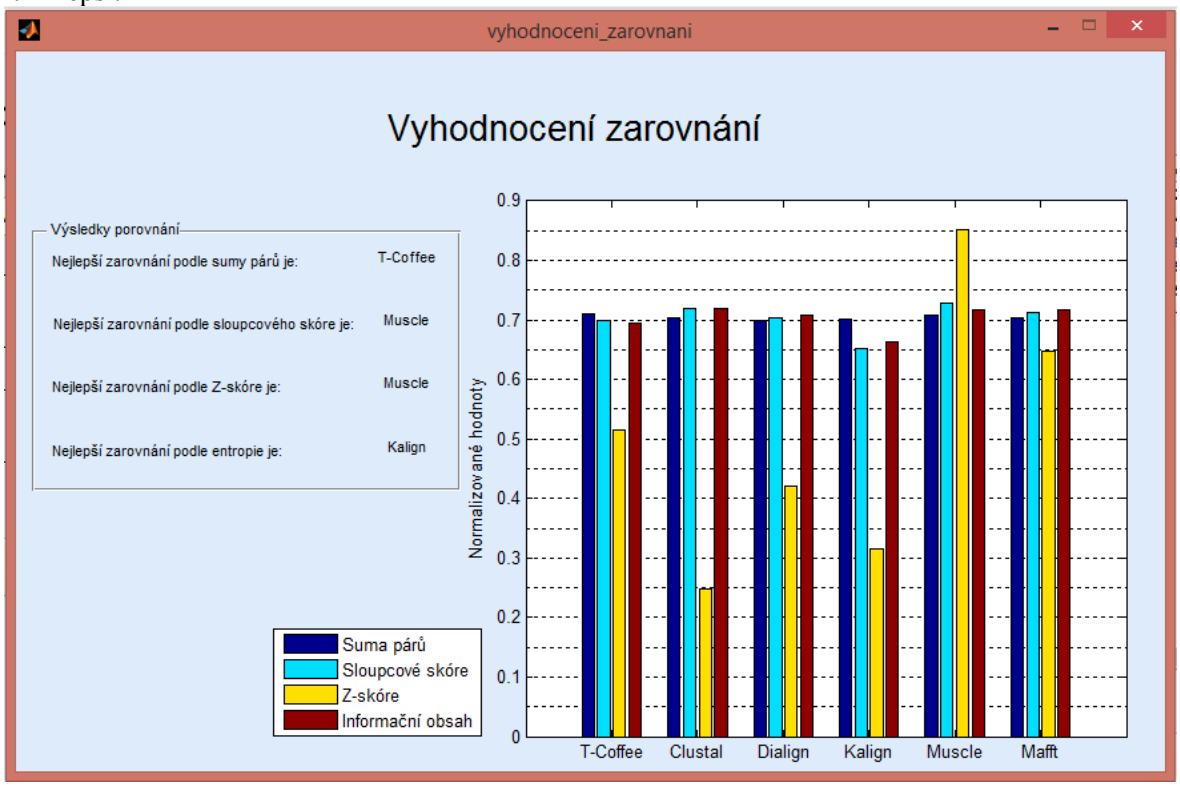


Obr. 5: Volba typu porovnání

V prvním případě je důležité výše zmíněné pojmenování. Toto porovnání je dále již pasivní, uživatli se zobrazí okno s grafickým znázorněním získaných hodnot a jednoduchým vyhodnocením dle parametrů jak je vidět na Obr. 6. Je možné si všimnout, že v grafu není uvedena entropie. Místo ní je zvoleno zobrazení informačního obsahu, která lze z entropie přepočítat pomocí vzorce (1)

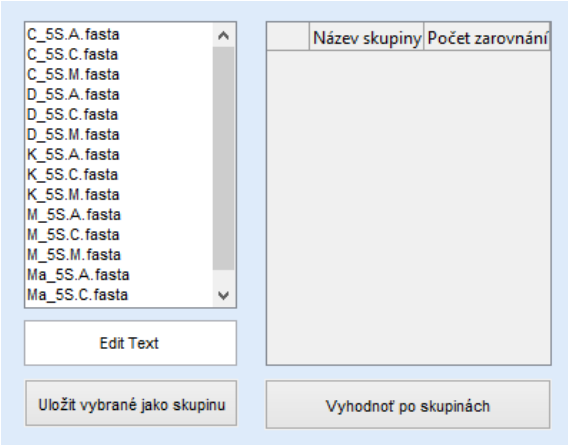
$$io = \frac{\log_2(d) - entropie}{\log_2(d)}, \tag{1}$$

kde d je 4 pro nukleotidy a 24 pro aminokyseliny.
Výhodou a důvodem této volby je, že můžeme všechny parametry posuzovat, čím vyšší hodnota, tím lepší.



Obr. 6: Vyhodnocení pro jedno zarovnání z různých nástrojů

V druhém případě je třeba dále aktivně postutovat, k porovnání je třeba zvolit skupiny (např. z jednoho nástroje, jednoho nastavení aj.). Zobrazeno na Obr. 7.



Obr. 7: Volba skupin

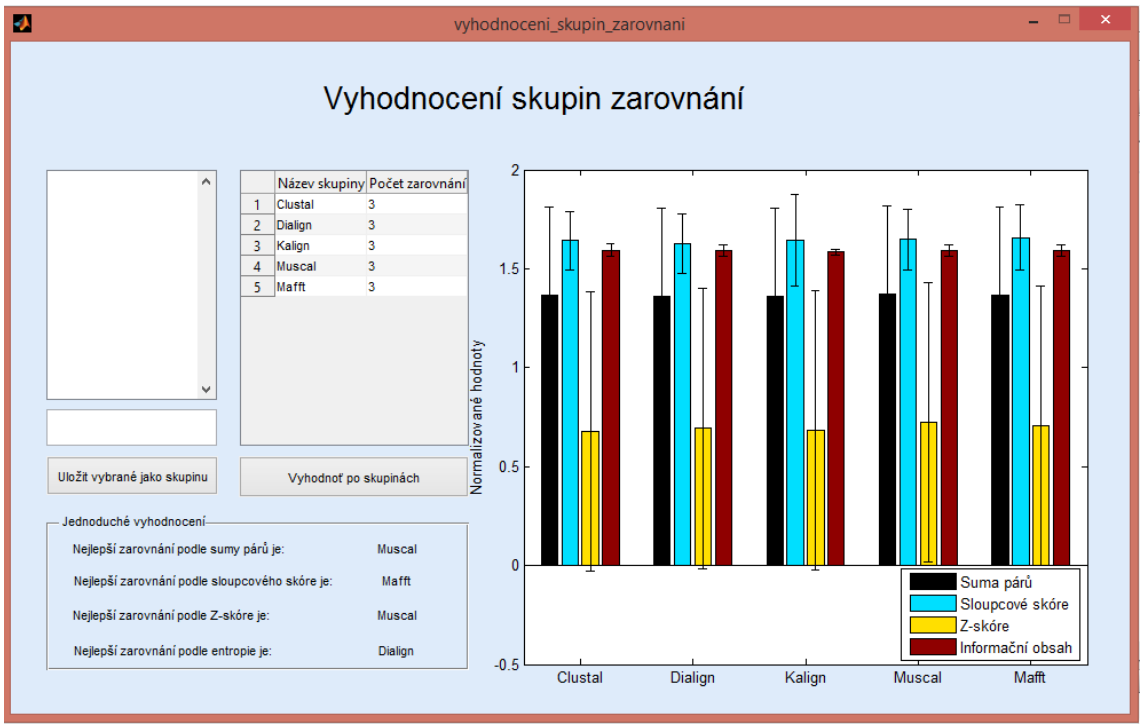
Následuje jednoduché vyhodnocení na základě průměru získaných normalizovaných hodnot. Normalizace je provedena matlabovskou funkcí normc s přičtením 1 a podělením 2. Na grafu jsou znázorněné průměry a ukázané rozpětí je směrodatná odchylka.

Celá normalizace je provedena následovně:
Pro daný sloupec x_i o délce r je vypočtená hodnota n

$$n = \frac{1}{\sqrt{(x_i * x_i) + 1}} \tag{2}$$

Protože x_i je vektor, tak pomocí vektorového násobení získáme jednu hodnotu.
Dále každou jednu hodnotu získáme následovně:

$$x_i(r) = \frac{[x_i(r) * n] + 1}{2} \tag{3}$$



Obr. 8: Porovnání zarovnání

Závěr

Program Nástroj_testování_kvality.m je sestaven pro porovnávání vícenásobného zarovnání. Program uživateli nabízí dvě formy porovnání, podle toho, jaký je charakter zarovnání. Uživatel má možnost hodnocení provést tak jak zarovnání je, nebo si nechat sestavit bloky, které budou následně ohodnoceny, tím nebrat v potaz např. dorovnání jedné dlouhé sekvence mezerami v ostatních sekvencích. Je možné ovlivnit ohodnocení změnou různých parametrů a tím přizpůsobit své potřeby. Kromě výpočtu parametrů je možné zobrazit graf a jednoduché doporučení pro daná zarovnání s ohledem na jednotlivé parametry.